# LONG-READ RNA SEQUENCING ANALYSIS OF THE LYTIC HUMAN CYTOMEGALOVIRUS TRANSCRIPTOME

## Ph.D. Thesis

**DR. ZSOLT BALÁZS**

Department of Medical Biology

Doctoral School of Multidisciplinary Medicine

Faculty of Medicine

University of Szeged

Supervisor: Prof. Zsolt Boldogkői, PhD, DSc

Szeged

2019

*Dedicated to my parents for their constant support during the many long years of my education.*

# 1 List of publications:

## 1.1 Publications directly related to the subject of the thesis:

I. Balázs, Zsolt; Tombácz, Dóra; Szűcs, Attila; Michael, Snyder; Boldogkői, Zsolt
Dual platform long-read RNA-sequencing dataset of the human cytomegalovirus lytic transcriptome
*FRONTIERS IN GENETICS* 9 Paper: 10.3389/fgene.2018.00432 (2018)
**IF: 4.151**

II. Tombácz, Dóra; Balázs, Zsolt; Csabai, Zsolt; Michael, Snyder; Boldogkői, Zsolt
Long-Read Sequencing Revealed an Extensive Transcript Complexity in Herpesviruses
*FRONTIERS IN GENETICS* 9 Paper: 259 (2018)
**IF: 4.151**

III. Balázs, Zsolt†; Tombácz, Dóra†; Szűcs, Attila; Michael, Snyder; Boldogkői, Zsolt
Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform
*SCIENTIFIC DATA* 4 Paper: 170194 (2017)
**IF: 5.305**

IV. Balázs, Zsolt; Tombácz, Dóra; Szűcs, Attila; Csabai, Zsolt; Megyeri, Klára; Alexey, N Petrov; Michael, Snyder; Boldogkői, Zsolt
Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials
*SCIENTIFIC REPORTS* 7 Paper: 15989, 9 p. (2017)
**IF: 4.122**

## 1.2 Publications indirectly related to the subject of the thesis:

V. Boldogkői, Zsolt*; Balázs, Zsolt; Moldován, Norbert; Prazsák, István; Tombácz, Dóra
Novel Classes of Replication-associated Transcripts Discovered in Viruses
*RNA BIOLOGY* 1 Paper 1-10 (2019)
**IF: 5.216**

VI. Boldogkői, Zsolt*; Szűcs, Attila; Balázs, Zsolt; Sharon, Donald; Snyder, Michael; Tombácz, Dóra*
Transcriptomic study of Herpes simplex virus type-1 using full-length sequencing techniques
*SCIENTIFIC DATA* 5 Paper: 180266 (2018)
**IF: 5.305**

VII. Prazsák, István†; Moldován, Norbert†; Balázs, Zsolt; Tombácz, Dóra; Megyeri, Klára; Szűcs, Attila; Csabai, Zsolt; Boldogkői, Zsolt*
Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus
*BMC GENOMICS* 19: 1 Paper: 873 (2018)
**IF: 3.730**

VIII. Moldovan, Norbert; Tombacz, Dora; Szucs, Attila; Csabai, Zsolt; Balazs, Zsolt; Kis, Emese; Molnar, Judit; Boldogkoi, Zsolt*
Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus.
*SCIENTIFIC REPORTS* 8: 1 p. 8604 (2018)
**IF: 4.122**

IX. Moldován, Norbert; Szűcs, Attila; Tombácz, Dóra; Balázs, Zsolt; Csabai, Zsolt; Michael, Snyder; Boldogkői, Zsolt*
Multi-platform Next-generation Sequencing Identifies Novel RNA Molecules and Transcript Isoforms of the Endogenous Retrovirus Isolated from Cultured Cells
*FEMS MICROBIOLOGY LETTERS* 365: 5 Paper: fny013, 6 p. (2018)
**IF: 1.735**

X. Moldován, Norbert; Balázs, Zsolt; Tombácz, Dóra; Csabai, Zsolt; Szűcs, Attila; Michael, Snyder; Boldogkői, Zsolt*
Multi-platform Analysis Reveals a Complex Transcriptome Architecture of a Circovirus
*VIRUS RESEARCH* 237 pp. 37-46., 10 p. (2017)
**IF: 2.484**

XI. Tombácz, Dóra; Csabai, Zsolt; Szűcs, Attila; Balázs, Zsolt; Moldován, Norbert; Donald, Sharon; Michael, Snyder; Boldogkői, Zsolt*
Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1
*FRONTIERS IN MICROBIOLOGY* 8 Paper: 1079, 16 p. (2017)
**IF: 4.019**

XII. Tombácz, Dóra†; Balázs, Zsolt†; Csabai, Zsolt; Moldován, Norbert; Szűcs, Attila; Donald, Sharon; Michael, Snyder; Boldogkői, Zsolt*
Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing
*SCIENTIFIC REPORTS* 7 Paper: 73751, 13 p. (2017)
**IF: 4.122**

XIII. Tombácz, D; Csabai, Z; Oláh, P; Balázs, Z; Likó, I; Zsigmond, L; Sharon, D; Snyder, M; Boldogkői, Z*
Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus
*PLOS ONE* 11: 9 Paper: e0162868, 29 p. (2016)
**IF: 2.806**

### 1.3 Publications not related to the subject of the thesis:

XIV. Tombácz, Dóra; Maróti, Zoltán; Kalmár, Tibor; Csabai, Zsolt; Balázs, Zsolt; Takahashi, Shinichi; Palkovits, Miklós; Snyder, Michael; Boldogkői, Zsolt*
High-Coverage Whole-Exome Sequencing Identifies Candidate Genes for Suicide in Victims with Major Depressive Disorder
*SCIENTIFIC REPORTS* 7 Paper: 7106, 11 p. (2017)
**IF: 4.122**

XV. Tombácz, Dóra; Moldován, Norbert; Balázs, Zsolt; Csabai, Zsolt; Michael, Snyder; Boldogkői, Zsolt*
Genetic Adaptation of Porcine Circovirus Type 1 to Cultured Porcine Kidney Cells Revealed by Single-Molecule Long-Read Sequencing Technology
*GENOME ANNOUNCEMENTS* 5: 5 Paper: e01539-16, 2 p. (2017)
**IF: 0**

**Cumulative impact factor: 55.390**

The 2018 impact factor values were not available at the time of writing, therefore the 2017 values are displayed for all publications newer than 2017.

# 2 Preface

The following work presents our group's efforts in the mapping of the human cytomegalovirus transcriptome. However, I would like to emphasize that these efforts meant so much more than merely cataloguing RNA isoforms of a pathogen. Our group was fortunate enough to have used state-of-the-art technology in a field that is barely beginning to flourish at the time of writing this manuscript. Long-read RNA sequencing presented multiple unexpected challenges both at the wet lab and at the informatics level. I must say it was quite uplifting as well as scary to walk an unwalked path. At the same time, it was inspiring to see the many innovative solutions emerging and I hope that our struggles contributed to paving the path for other researchers who decide to apply third-generation sequencing.

The manuscript is based on a series of RNA sequencing experiments carried out using two long-read sequencing technologies. The results of the analyses were published in multiple papers, of which the one about sequencing on the PacBio RSII platform was already featured in the thesis of my friend and colleague, Dr. Zsolt Csabai, who presented the utility of RNA sequencing in a multitude of herpesviruses. Therefore, those findings will not be the main focus of this thesis. However, as those findings are relevant to the subject of the current thesis, they will be explained and referenced to contrast them to results obtained by different platforms. In order to show a more complete picture of the projects of our group and to provide a better narrative, some as of yet unpublished data is presented.

# 3 Table of contents

# 4 Introduction

## 4.1 Sequencing and genome annotation

With the rapid fall of the cost of sequencing the field of genomics has expanded rapidly. The sequencing of large genomes such as the human genome seemed an impossible task 30 years ago and still a very expensive one 20 years ago (International Human Genome Sequencing Consortium, 2001). In the last decade, however, next-generation sequencing (NGS) has made it possible to sequence a human genome in days for the cost of a cell phone (Goodwin et al., 2016). With the use of NGS, many fields were able to answer interesting and essential questions. NGS has been used to uncover the history of the human race and also of more recent history of smaller populations (Green et al., 2010; Lazaridis et al., 2014). NGS has been used to investigate cancer and aging (Campbell et al., 2008; Lodato et al., 2018). And it has revealed that our body hosts a wide range of organisms (Moustafa et al., 2017; Turnbaugh et al., 2007). Not only do we know the genome sequence of numerous organisms (Adams et al., 2000; Bult et al., 2007; Schnable et al., 2009), but we can compare the genetic variability of thousands of individuals (Turnbull et al., 2018). Although we possess a lot of genome sequences, we only know the function of a very small fragment of these sequences.

Genome annotation is the process of dividing the genome into functional regions and determining the functions of these units. This can involve gene prediction or other structural annotation of larger genetic elements, the mapping of binding sites for various proteins or the functional characterization of variants (Harrow et al., 2012).

One of the many parts of genome annotation is the annotation of the transcriptome. RNAs are the molecules that relay the information coded in the DNA outside of the nucleus to determine the amino acid sequence of proteins, but RNAs can also act as enzymes and also possess a number of other regulatory roles (Mercer et al., 2009). Transcriptome annotation means the mapping of transcriptionally active genetic regions, the structural characterization of RNA molecules and the functional categorization of the transcripts. A complete and detailed transcript annotation contributes to the interpretation of genetic variants as well as to the analysis of quantitative RNA sequencing (RNA-Seq) data (Soneson et al., 2015; Yu et al., 2007). When analyzing genetic mutations, it is crucial to know whether a certain variant is

transcribed and if it is, then whether it is in an untranslated region or whether it is alternatively spliced etc. Also, in the analysis of gene expression, it is essential to know how many and what kind of transcripts are expressed from a region, because it can greatly improve the accuracy of the programs which (Trapnell et al., 2012) analyze RNA-Seq data.

Genes can express a variety of transcripts. The processes that contribute to alternative gene usage are alternative polyadenylation (Edwalds-Gilbert et al., 1997; Tian et al., 2005), alternative splicing (Roca et al., 2003), alternative promoter usage (Cramer et al., 1997) and RNA base modifications (Hussain et al., 2013).

### 4.2 NGS revolutionizes transcriptomics

Compared to previous tools that were available for transcript analysis such as reverse-transcription quantitative PCR and microarrays, RNA-Seq was a rapid advancement. NGS enabled the simultaneous investigation of multiple genes without the need for prior knowledge about the sequence. As the price of sequencing fell rapidly, more and more transcriptomes and complete transcriptome atlases of whole human organs became available (Hawrylycz et al., 2012). Soon, more specific applications were developed for the more detailed characterization of transcriptomes. PolyA-Seq emerged to characterize the 3’ ends of transcripts and to discover alternative polyadenylation (Shepard et al., 2011). Sequencing of the 3’ ends of the transcripts proved to be especially beneficial, as RNA degradation generally proceeds from the 5’ end. The cap analysis of gene expression (CAGE) was developed to map the 5’ ends of the transcripts (Kodzius et al., 2006). Novel variations of NGS such as GRO-Seq and PRO-Seq are used for the analysis of nascent RNA as it is being transcribed (Gardini, 2017; Mahat et al., 2016). Short-read sequencing can efficiently characterize transcript features such as transcriptional start sites (TSS), transcriptional end sites (TES) and introns, however, when multiple transcripts overlap, short reads cannot tell, to which transcript a given read belongs.

### 4.3 The third-generation of sequencing.

Long-read sequencing is a novel tool in genomics and it is slowly gaining appreciation and it is being constantly improved in order to be able to applied to answer a variety of questions (van Dijk et al., 2018). Two companies dominate the field of long-read sequencing

at the moment: Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), which are providing technologies based on different principles of nucleotide detection.

### 4.3.1 Pacific Biosciences

Pacific Biosciences patented Single-Molecule Real-Time (SMRT) sequencing, which relies on a polymerase molecule fixated at the bottom of a zero-mode waveguide, a tiny structure that enables the detection of fluorescence emitted during the incorporation of a single labelled nucleotide. The fluorescence reaches a detector, which converts the signal. The template that the polymerase has to process is circularized using hairpin adapters (SMRTBell). The polymerase can make multiple rounds on each circularized template. The signals of multiple passes over the same sequence can be summed to create a more accurate circular consensus sequence (CCS). The disadvantages of PacBio sequencing are that it is less accurate and produces much less reads than current short-read sequencing approaches. Its advantages are that it can produce reads even longer than 10 kb, and that its errors are more random than those of any other currently available sequencing technology. This means that if a high enough number of passes is reached over a given sequence, its accuracy can be improved above any other sequencing method's accuracy (Rhoads and Au, 2015; Roberts et al., 2013). PacBio sequencing has been used to improve genome assemblies, taking advantage of the long reads, that it provides (English et al., 2012) and also to characterize structural variations which are difficult to discern from short-read sequencing data (Nakano et al., 2017). PacBio achieved great success with its RSII platform; its improved version, the Sequel platform (Hebert et al., 2018) has recently been released and has a substantially higher throughput.

### 4.3.2 Oxford Nanopore Technologies

Oxford Nanopore Technologies (ONT) was the first company to utilize nanopore sequencing commercially in nucleic acid sequencing (Wang et al., 2018). Nanopore sequencing is based on the electric potential changes that the DNA or RNA molecules generate when they pass through a nanopore. The measured changes serve as signals for basecalling. The accuracy of nanopore sequencing is even lower than that of PacBio sequencing (Weirather et al., 2017). However, its throughput is higher and also the maximum read length is independent of the technology, only DNA extraction and library preparation impose a limit to the size of the fragments that can be sequenced (Jain et al., 2018). The accuracy of the ONT

platform can be increased somewhat by the $1D^2$ method. This technology attaches special adapters to one end of the DNA, which increase the chance that a strand will be followed by its complementary strand through the nanopore. This allows for a consensus sequence to be called from twice as many signals, thereby improving the accuracy. The users of nanopore sequencing have come up with other techniques to improve the quality of the reads. The R2C2 method utilizes rolling circle replication to produce concatamers of the template, which are then used to create a consensus (Volden et al., 2018). Further advantages of this technology are that it only requires a very low amount of capital investment and that the sequencing device is small and portable, therefore it can be applied in the field and does not require a large laboratory. Nanopore sequencing has also been utilized for genome assembly and for the detection of structural variations (Goodwin et al., 2015; Norris et al., 2016). For genome sequencing purposes, nanopore sequencing is often accompanied by an NGS method to improve the accuracy (Madoui et al., 2015). Both ONT and PacBio sequencing can be used to detect base modifications.

### 4.3.3 Long-read RNA sequencing

Long-read sequencing has clear advantages over short-read sequencing in transcriptome analysis. The error-prone nature of long-reads generally poses little to no challenge if the RNA sequencing is carried out on an organism with a known genome. However, long-reads provide full contig information about the transcripts, which means that the transcript isoforms can be differentiated by TSS, TES and exon connectivity (**Figure 1**). RNA-Seq initially meant, and mostly still refers to, cDNA sequencing, that is, the RNA has to be reverse transcribed first into cDNA and then the cDNA is sequenced. Both long-read sequencing technologies take advantage of the switching mechanism at the 5’ end of the RNA transcript (SMART), which enables the second strand synthesis of full-length cDNA molecules (Gonzalez-Garay, 2016; Zhu et al., 2001). Even though this method is widely used and accepted, its results are difficult to interpret due to effects of RNA degradation and technical artefacts.

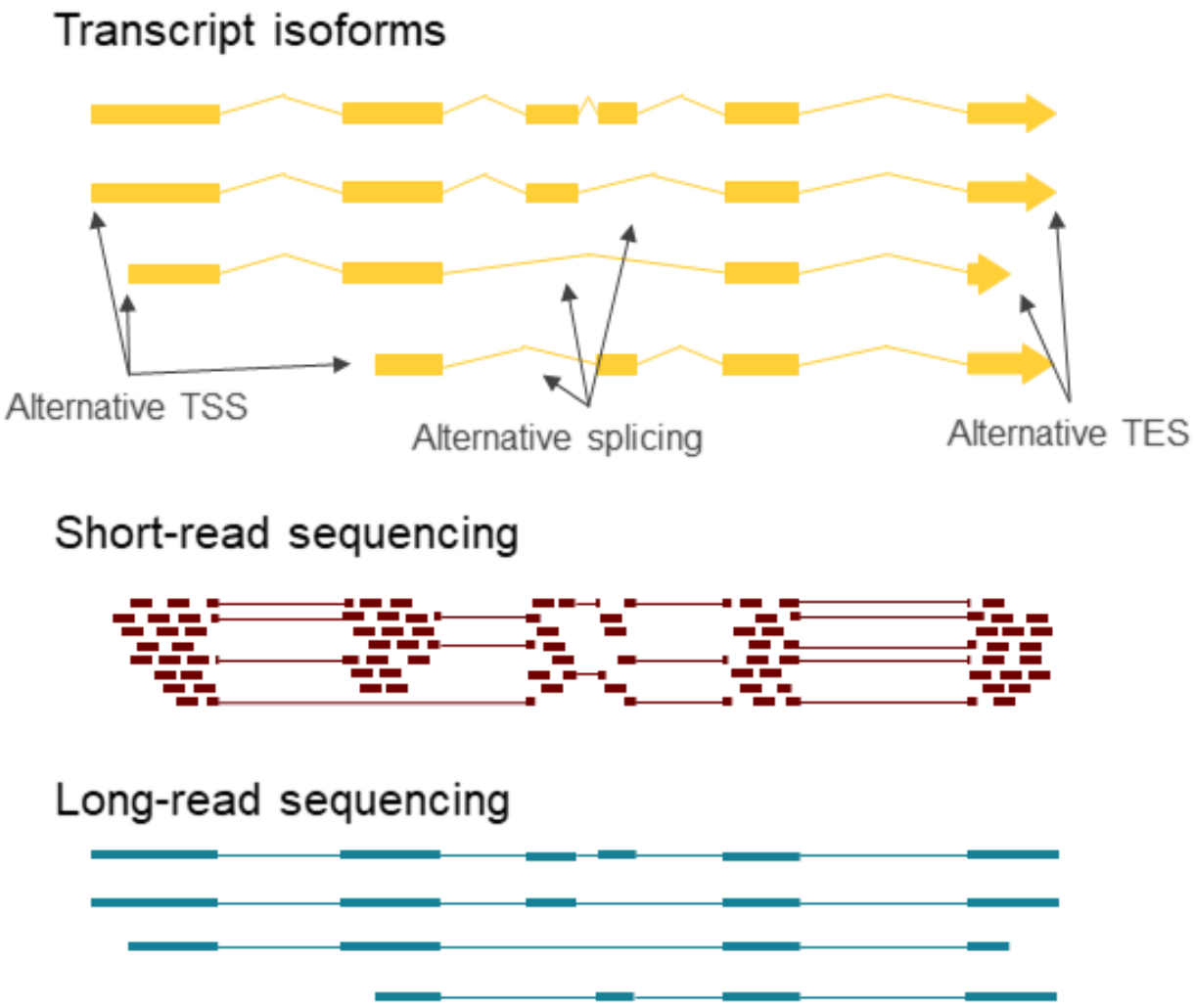


***Figure 1 The advantages of long-read sequencing.*** *The yellow rectangles represent the transcribed exons of a gene, red rectangles represent short reads, blue ones represent long reads. The thin lines represent introns. It is visible, that the different transcripts could not be reconstructed using short-read sequencing, while long-read sequencing is able to differentiate the transcript isoforms.*

### 4.4 The drawbacks of cDNA sequencing

Reverse transcription and PCR are known to be able to produce artefacts through template switching (Geiszt et al., 2004; Mader et al., 2001; Zeng and Wang, 2002). Template switching is the process through which the polymerase stops elongation and reinitiates at a different locus, which contains homologous sequences. Reinitiation is thought to be facilitated by the ability of the polymerase to bind to the synthesized strand even after the polymerase has dissociated from the template. The strand is carried over and hybridizes to a homologous locus

from where the polymerase can continue elongation. The DNA polymerase and the reverse transcriptase are both capable of template switching (Cocquet et al., 2006; Kanagawa, 2003). Template switching has been described to introduce chimeric cDNAs (Brakenhoff et al., 1991), which may resemble splicing or trans-splicing events (Cocquet et al., 2006; Roy and Irimia, 2008), and may also appear as antisense transcripts (Yuan et al., 2013). Template switching can even occur between as short as 3 nucleotide-long homologous sequences, although generally, longer homologous sequences cause template switching more frequently (Dang and Hu, 2001). Increasing the concentration of the templates or lowering the temperature during reverse transcription may also increase the frequency of template switching.

Another common technical artefact in cDNA sequencing is internal priming. Internal priming is the ability of 5' or 3' end adapters (such as the oligod(T) primer) to serve as primers at homologous positions inside the transcript, thereby truncating the transcript. Internal priming has been known to produce many artefacts in expressed sequence tags and in polyA-Seq (Nam et al., 2002). Several measures can be taken in order to reduce the effects of internal priming. Adenine-rich genomic regions can be excluded from the analysis of polyadenylation by bioinformatic filtering. Usually, sites which contain six or more consecutive adenines are discarded (Aaronson et al., 1996; Gautheret et al., 1998). Another solution is to devise library preparation methods which are immune to internal priming artefacts, such as the 3' READS+ method (Zheng et al., 2016).

### 4.5 Direct RNA sequencing

Direct RNA (dRNA) sequencing was first developed by Helicos as a short-read sequencing method (Ozsolak et al., 2009). However, as the company went bankrupt, the only commercially available platform, at the time of writing, is ONT (Garalde et al., 2018). The same device that sequences DNA, is also able to sequence RNA using the same nanopore technology. The advantage of dRNA sequencing is that it is not affected by the artefacts of reverse transcription or PCR. Both sequencing technologies can sequence non-amplified cDNA molecules, which are also free from PCR bias and currently produce a higher throughput. However, dRNA sequencing can also detect base modifications in native RNA molecules, thereby revolutionizing the field of epitranscriptomics (Li et al., 2017).

### 4.6 The bioinformatics of long-read RNA sequencing

The data generated during long-read sequencing is different from short-read sequencing data, therefore different bioinformatic tools are needed for their analysis. Only few tools exist for the processing of long-read sequencing reads, and usually these tools are very new, consequently, no standardized method has been devised yet for the handling of long reads. The preprocessing of the data is different for the two technologies; SMRT Analysis is the software suite that is used for the primary analysis of PacBio reads and MinKNOW is the one for ONT reads. The two commonly used aligners are GMAP (Wu and Watanabe, 2005) and minimap2 (Li, 2018), and they function equally well with reads provided by both technologies. The downstream analysis of the reads is less standardized with some programs being published during the writing of this manuscript. PacBio reads can be assembled into IsoSeq consensus isoforms by the SMRT Analysis software. These isoforms depict full-length transcripts and can be used for downstream analysis. However, a quality filtering of these isoforms is advised; SQANTI characterizes the isoforms detected by long-read sequencing and also filters them based on their quality (Tardaguila et al., 2018). Tools for the clustering of nanopore reads into consensus isoforms are also emerging, and are generally based on concepts similar to the IsoSeq protocol of PacBio (https://github.com/nanoporetech/pinfish, https://github.com/BrooksLabUCSC/flair).

### 4.7 The human cytomegalovirus

The human cytomegalovirus (HCMV) is a human pathogen betaherpesvirus which has a wide range of seroprevalence, depending on age, geographic location and demographics (Cannon et al., 2010). The infection generally has mild symptoms or can be asymptomatic in healthy adults and children. Sometimes the infection causes mononucleosis-like symptoms (Vancíková and Dvorák, 2001). HCMV can establish latency in CD33+ or CD34+ progenitor cells, neutrophil granulocytes or monocytes (Schottstedt et al., 2010). Reactivation or new infection during pregnancy often lead to severe complications (Davis et al., 2017; Wen et al., 2002). HCMV infection has been detected in almost all cases of glioblastoma, and a high-grade viral infection correlated with worse prognosis (Rahbar et al., 2013).

The virus has a linear dsDNA genome of 235 kb, which is the largest genome in the *Herpesviridae* family. The genome consists of two unique regions (unique long: UL and unique short US) each flanked by inverted repeat sequences. It is estimated that HCMV expresses between 164 and 220 proteins (Davison et al., 2003; Murphy et al., 2003). It is difficult to maintain clinical isolates in cell cultures. The highly passaged strains which adapted to cell cultures usually contain multiple mutations (Dolan et al., 2004).

HCMV replicates in the nucleus. Upon arrival in the nucleus, the tegument proteins of the virus serve as transactivators and recruit the host RNA polymerase II to transcribe the immediate-early genes of HCMV. The immediate-early genes are transactivator proteins that initiate the transcription of early genes and also regulate the expression of the host MHC I proteins (Schottstedt et al., 2010). The early genes are required for the viral DNA synthesis. Following the start of DNA synthesis, the late genes are expressed, which code for the structural proteins; they are necessary for the assembly and egress of the virus (Fields et al., 2013).

Similarly to other herpesviruses, HCMV has a rather complex transcriptional architecture: its genes are often arranged into large, overlapping polycistronic clusters (Ma et al., 2012). In contrast to alphaherpesviruses, betaherpesviruses such as HCMV, utilize many splice junctions (Gao et al., 2015; Gatherer et al., 2011; Rawlinson and Barrell, 1993). Alternative transcription initiation allows for the expression of multiple different proteins from the same gene (Arend et al., 2016; Caviness et al., 2014; Isomura et al., 2008).

Up until recently, the HCMV transcriptome research used Northern blotting, RT-qPCR and Rapid amplification of cDNA ends techniques (He et al., 2012; Kondo et al., 1996; Ma et al., 2011). These methods can only analyze one gene at a time and the mapping of the whole HCMV transcriptome would prove to be a tedious work. The HCMV transcriptome was first analyzed using NGS in 2011. The study discovered that alternative splicing is very common in HCMV, and that a large part of the genome expresses antisense transcripts (Gatherer et al., 2011). The study also found that five long non-coding intergenic RNAs are responsible for almost two thirds of the HCMV reads. Another RNA-Seq study examined the translational capacities of the virus (Stern-Ginossar et al., 2012). Ribosome profiling (the sequencing of mRNA stretches that are covered by ribosomes) had defined 751 translationally active open

reading frames (ORF) in the HCMV genome. Many of these ORFs were coding for short oligopeptides and were positioned upstream of the main proteins. Another fraction of those proteins were N-terminally truncated ORFs or the main proteins. The study has also found that some HCMV ORFs are not initiated with an AUG (methionine), but instead with other triplets, which may or may not show similarity to the Kozak sequence (Stern-Ginossar et al., 2012). A more recent RNA sequencing study used single-cell sequencing to examine the HCMV transcriptome during latency, and found that in contrast to most other herpesviruses, HCMV expresses no specialized latency transcript, instead the same transcripts are expressed as during lytic infection, albeit in different quantities (Shnayder et al., 2018).

# 5 Aims

Long-read RNA sequencing has effectively characterized the transcriptome of several organisms before (Abdel-Ghany et al., 2016; Moldován et al., 2018; O'Grady et al., 2016; Prazsák et al., 2018; Sharon et al., 2013; Tombácz et al., 2016, 2017b; Wang et al., 2016; Workman et al., 2018). In all of the cases, long-read sequencing has revealed countless novel transcripts and transcript isoforms (Tombácz et al., 2018).

Although many studies have investigated the HCMV transcriptome before, most of its genes were not transcriptionally annotated. In order to characterize the lytic transcriptome of HCMV, we sequenced RNA isolated from HCMV-infected human fibroblast cells. We have used several different long-read sequencing library preparation methods and sequenced those libraries on different long-read sequencing platforms. Our aim was to draw a detailed map of the HCMV transcriptome.

Further, we planned on comparing the used sequencing methods in terms of performance and accuracy. We have used dRNA sequencing in order to be able to differentiate technical artefacts from real transcripts.

In order to be able to automate the analysis of the data and also to be able to compare the results obtained by the different sequencing platform and sequencing methods, we needed to develop a pipeline for the analysis of long-read RNA sequencing data.

# 6 Materials and methods

## 6.1 Samples

The data stem from the examination of two different biological samples (with Biosample accession numbers ERS1870077 and ERS2312967). A summary of the layout of the experiments is shown in ***Figure 2***.

### 6.1.1 Biosample ERS2312967

Four T75 cell culture flasks of MRC-5 cells [embryonic human lung fibroblast; American Type Culture Collection (ATCC) CCL-171] were cultured at 37°C and 5% $CO_2$-concentration in Dulbecco's Modified Eagle Medium supplemented with 10% fetal bovine serum and 100 units of potassium penicillin and 100µg of streptomycin sulphate per 1ml. Rapidly-growing near-saturated cell cultures were infected with the strain Towne of HCMV (ATCC VR-977), at a multiplicity of infection of 0.5 plaque-forming units per cell. The infected cells were incubated for 1h, then the virus suspension was removed and washed with phosphate-buffered saline. Subsequently, the cells were incubated in fresh culture medium for 24, 72 or 120h (in 1-2-1 flasks respectively). Total RNA was isolated using the NucleoSpin RNA kit and aliquots of 20µl from each sample were pooled before reverse transcription.

### 6.1.2 Biosample ERS1870077

Eight flasks of MRC-5 cells were infected with HCMV strain Towne VarS (ATCC VR-977) and incubated under the same conditions as described above. A virus titer of 0.05 plaque forming units per cell was used for infection. Total RNA was isolated from the infected cells at 1h, 3h, 6h, 12h, 24h, 72h, 96h, 120h post infection.

## 6.2 Selection and library preparation

Polyadenylated RNAs were selected from both samples using Oligotex mRNA Mini Kit. Five different, poly(A)-selected libraries were prepared in order to better characterize the HCMV transcriptome.

### 6.2.1 Poly(A)+, not cap-selected cDNA library for sequencing on the ONT platform

31 ng polyA(+) RNA from biosample ERS1870077 was used for first strand cDNA synthesis using SuperScript IV and adapter-linked oligod(T) primers, then 5' adapter sequences

with three O-methyl-guanine ribonucleotides were ligated to carry out second-strand synthesis. The cDNA was amplified through 18 cycles using KapaHiFi DNA polymerase. UltraPure Agarose gel was used to separate PCR products and the cDNA fragments larger than 500 nucleotides were isolated using the Zymoclean Large Fragment DNA Recovery Kit. The library was prepared using the Ligation Sequencing 1D kit (SQK-LSK108) and the NEBNext End repair/dA-tailing Module NEB Blunt/TA Ligase Master Mix according to the manufacturers' recommendations. The library was sequenced on one MinION flowcell.

### 6.2.2 Poly(A)+, cap-selected cDNA library for sequencing on the ONT platform

2 µg of total RNA of biosample ERS2312967 was reverse transcribed using the TeloPrime Full-Length cDNA Amplification Kit. The TeloPrime protocol contains both poly(A) and cap selection- The 5' adapter was ligated to the DNA-RNA hybrid overnight at 25 °C. Endpoint PCR (of 30 cycles) was performed using the reagents supplied in the kit. The sequencing library was prepared using the Ligation Sequencing 1D kit (SQK-LSK108) and the NEBNext End repair / dA-tailing Module NEB Blunt/TA Ligase Master Mix according to the manufacturers' recommendations. One MinION flowcell was used for the sequencing.

### 6.2.3 Poly(A)+ Direct RNA library for sequencing on the ONT platform

500 ng polyA(+) RNA was used from biosample ERS2312967 for direct RNA sequencing. A first-strand cDNA was synthesized using SuperScript IV (Thermo Fischer Scientific) and the adapter primers provided by the Direct RNA Sequencing kit (SQK-RNA001). The library was prepared following the instructions of the manufacturer. The library was sequenced on one MinION flowcell.

### 6.2.4 Poly(A)+, cDNA library for sequencing on the RSII platform

2 ng polyadenylated RNA from biosample ERS1870077 was reverse transcribed using SuperScript IV and anchored oligod(T) primers, following the PacBio Iso-Seq protocol. A PCR of 18 cycles was performed by using the Clontech SMARTer PCR kit to amplify the cDNA. No size selection was performed on the sample. The library preparation was carried out using SMRTbell DNA Template Prep Kit 2.0 and the MagBead Kit. P6-C4 chemistry was used for the sequencing. 7 SMRT cells were used to sequence the library.

### 6.2.5 Poly(A)+, cDNA library for sequencing on the Sequel platform

2 ng poly(A)-selected RNA from biosample ERS1870077 was reverse transcribed using SuperScript IV and anchored oligod(T) primers, following the PacBio Iso-Seq protocol. The cDNA was amplified using the Clontech SMARTer PCR (18 cycles). The cDNA sample was not fractionated according to size. The library was prepared with the SMRTbell DNA Template Prep Kit 2.0 and bound to MagBeads (MagBead Kit v2) for sequencing using the P6-C4 chemistry on one SMRT cell.

### 6.2.6 Random-primer cDNA library for sequencing on the RSII platform

Total RNA from biosample ERS1870077 was depleted of rRNA using the RiboMinus™ Eukaryote System v2 kit. 2ng of the remaining RNA was used for random primer driven cDNA synthesis. The following steps were the same as described for the poly(A)+ cDNA that was sequenced on the RSII platform. The library was sequenced on one SMRT cell.

### 6.2.7 Not cap-selected, random-primer cDNA library for sequencing on the ONT platform

2 µg total RNA from biosample ERS1870077 was treated with RiboMinus™ Eukaryote System v2 kit. 60ng of the remaining RNA was reverse transcribed using random primers. The following steps were the same as described for the not cap-selected poly(A)+ cDNA library sequenced on the ONT platform.

### 6.2.8 Cap-selected, random-primer cDNA library for sequencing on the ONT platform

2 µg total RNA from biosample ERS1870077 was reverse transcribed using the TeloPrime Full-Length cDNA Amplification Kit with random primers instead of oligod(T) primers. No ribodepletion was performed. The following steps were the same as described for the cap-selected poly(A)+ cDNA library sequenced on the ONT platform. The cap-selected and the not cap-selected random cDNA libraries were sequenced on the same MinION flowcell using barcodes supplied in the ONT Barcoding PCR kit.

### 6.2.9 Technical validation

Sample concentration was determined using the Qubit (ds)DNA HS Assay Kit. PacBio's Binding Calculator in RS Remote was used to determine the conditions for primer annealing and binding of the polymerase when loading the PacBio samples. An Agilent 2100 bioanalyzer was used to measure the concentrations of the PacBio libraries.

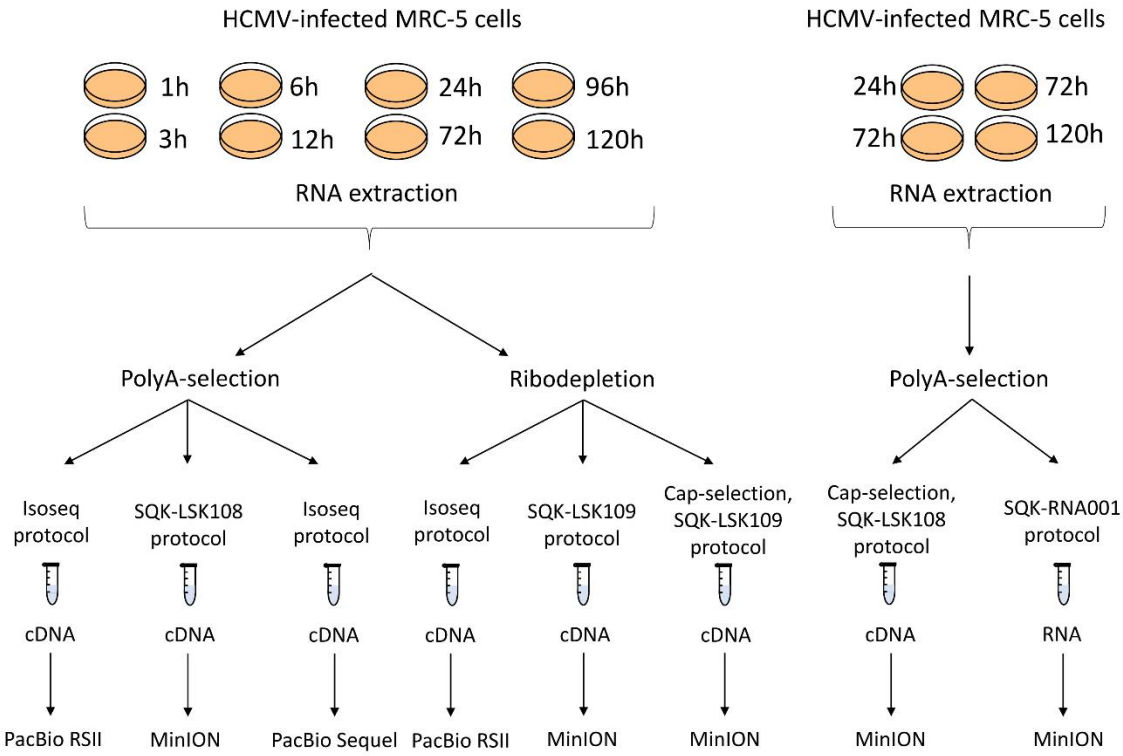


***Figure 2 Experimental layout.*** *MRC-5 cells were infected with the Towne strain of HCMV. The isolated RNA samples from different post infection time points were pooled into two independent biosamples (accessions: ERS1870077 and ERS2312967). Total RNA samples were subjected to either ribodepletion (random cDNA libraries) or to poly(A)-selection.*

## 6.3 Sequencing

### 6.3.1 MinION platform

All five libraries were sequenced on R9.4 SpotON Flow Cells with a MinION DNA/RNA sequencing device. The sequencing runs were carried out using MinKNOW. Voltage levels were set and reset in line with the suppliers' recommendations. Base calling was performed using Albacore v1.2.6.

### 6.3.2 RSII platform

The two libraries (altogether eight SMRTcells) were sequenced on a single SMRT cell using the RSII system. The length of the run was 4 h. Consensus sequences were generated using SMRT Analysis version v2.3.0 (Potter), using the RS_ReadsOfInsert protocol.

### 6.3.3 Sequel platform

The prepared library was sequenced on a single SMRT cell using the Sequel system. The length of the run was 10 h. Consensus sequences were generated using SMRT-Link v5.0.1 (Potter).

## 6.4 Read preprocessing

Neither the nanopore, nor the PacBio reads were trimmed. No error-correction method was applied to the reads. All reads were mapped by GMAP (Wu and Watanabe, 2005) with the following arguments: gmap -d *Reference*.fa --nofails -f samse *Reads*.fastq > *Mapped_reads*.sam. The files were compressed to binary .bam files using samtools view (Li et al., 2009). Read statistics were extracted using custom scripts.

## 6.5 A pipeline for transcript discovery

The original pipeline that was used for transcript discovery during the analysis of the RSII sequencing data (Balázs et al., 2017a) is summarized in **Figure 3**. In brief, consensus reads were mapped to the HCMV genome. Reads with a high mismatch or indel ratio (>5%) were discarded. All good quality reads were used to identify introns. The deletions in reads that complied with the GT-AG rule (contained the sequences GT immediately downstream of the donor and AG immediately upstream of the acceptor sites) were accepted as splice junctions. Reads which contained at least 15 terminal (A) mismatches (i.e. a poly(A) tail) were considered for the validation of TESs. A TES was accepted as valid if two reads confirmed the same nucleotide position and the genomic region of the putative TES did not contain a stretch of 3 or more (A)s. Reads with a definite orientation were considered for the identification of TSSs. If the number of reads starting at a given genomic position was significantly higher than that would be expected according to the Poisson distribution, the genomic position was accepted as a TSS. Transcript isoforms were annotated based on reads containing the above-mentioned annotated features.

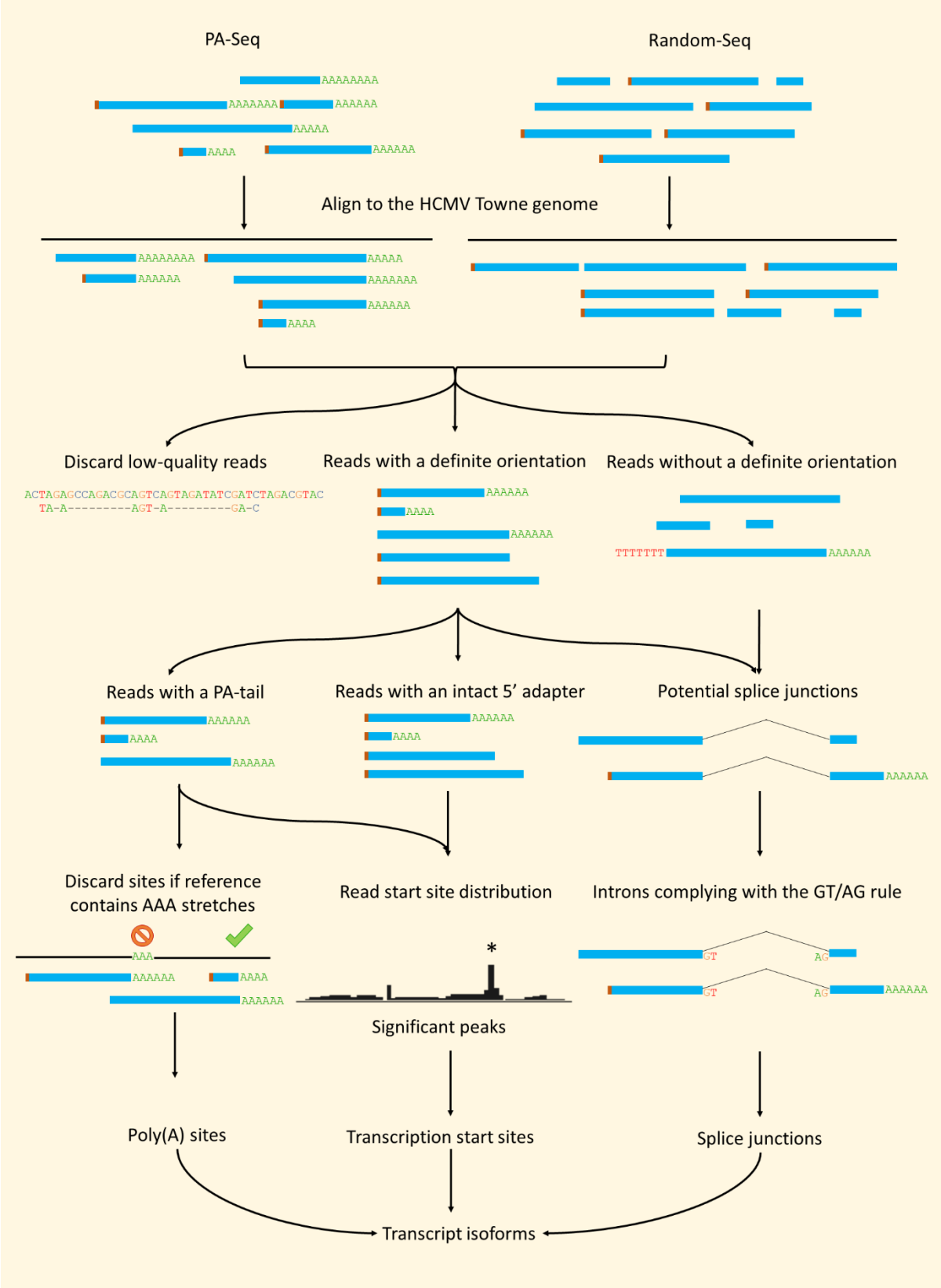


***Figure 3 The transcriptome profiling pipeline used for the RSII sequencing data.***

## 6.6 The LoRTIA toolkit

The original pipeline was significantly improved upon to streamline the analysis of long-read RNA sequencing data. The new Long-read RNA-Seq Transcript Annotator (LoRTIA) toolkit (**Figure 4**) is capable of handling a variety of long-read RNA sequencing inputs. The program is fast and modular and can therefore be used for multiple annotation purposes not merely to annotate transcriptomes. The source code is open and available online (Balázs, 2018). This toolkit was also used in the analysis of the varicella zoster transcriptome (Prazsák et al., 2018). In order to present a complete picture of the long-read RNA sequencing experiments of our group, all HCMV data have been processed by the LoRTIA software. The following paragraphs detail the functions of the software to show how the data were handled.

### 6.6.1 Accepted inputs

Accepted inputs are .bam or .sam files of aligned long-read sequences. The input reads should contain adapter sequences (untrimmed) otherwise the information on read orientation should be supplied separately (e.g. for IsoSeq sequencing or dRNA reads). Our recommendation is to use minimap2 (Li, 2018) as an aligner for all long-read data, however older datasets, which have been aligned by GMAP (Wu and Watanabe, 2005), are also accepted. Our dataset was aligned using the following command: minimap2 -ax splice -Y -C5 -t4 --cs *Reference*.fasta *Reads*.fastq > *Mapped_reads*.sam.

### 6.6.2 Software dependencies

The software is written in python 3 and runs in any UNIX environment. Biopython (Cock et al., 2009) is used for the reading and writing of sequence information. Pysam (Heger) is used for substituting samtools functions in python. Coverage data is obtained by bedtools (Quinlan and Hall, 2010). Data formatting and statistics are handled using the scipy and pandas packages (Jones et al., 2001).

### 6.6.3 Processing the input

The orientation of input reads (if not supplied otherwise) is determined based on the presence or absence of 5’ and 3’ adapter molecules. The 3’ adapter is the poly(A) tail during poly(A)-selected cDNA sequencing. The adapter sequences are searched in both orientations at both ends of each read. As long-read sequencing is notoriously error-prone, adapter-searches are not exact searches, the user can set the limit of similarity. The longer the adapter sequence,

the more sequencing error can be allowed. By default, adapter sequences are only searched in the 30 nucleotides upstream and 5 nucleotides downstream of the start of the alignment in the read. This increases specificity even when dealing with noisy data. If no adapter is detected or discordant adapters are detected on the two ends (e.g 3' adapters on both ends), the read direction is not specified, and the reads are tagged. If the end of the adapter sequence is detected 3 or more nucleotides downstream of the start of the alignment, the adapter sequence may have been placed there by template switching or internal priming, the program tags such reads as "potential template switching". While iterating over the reads, the program summarizes adapter statistics which are necessary for defining transcript features.

### 6.6.4 Detecting transcriptional start sites (TSS)

The distribution of read 5' ends for each nucleotide of the genome is determined, and local maxima (of ±10 nucleotide-size bins, created by sliding window mechanism), which account for at least 0.1% of the local coverage and are supported by at least two reads, are listed. The local maxima, where significantly more reads start than in its 101- nucleotide-long (±50 nucleotides) vicinity, are annotated as TSSs. The significance is determined based on the Poisson-probability (Poisson[$k_0$; $\lambda$]) of there being at least $k_0$ reads starting at a given nucleotide in 101-nt-long bin, where on average $\lambda$ reads ended.

### 6.6.5 Detecting transcriptional end sites (TES)

Every read that contains a poly(A) tail is used for the identification of TESs. Nucleotide positions which were supported by at least 0.1% (at minimum 2 reads) of the local reads, are accepted as TES. Only the local maximum of a ±10 nucleotide-size bin is considered a poly(A) site, in the cases of equal peaks in a bin, the most downstream position is selected.

### 6.6.6 Detecting introns

All reads which are mapped with an intron, bordered by exons of at least 25 nucleotides on both sides are used for the detection of splice sites. The mapped introns which are detected in at least two reads and in at least the 0.1% of the reads covering that genomic region, and also contain consensus splice site sequences (GT/AG, GC/AG or AT/AC) are accepted as introns. As indel errors are common in long-read sequencing platforms, rare deletions neighboring (not further than 20 nucleotides away from) frequent introns, are discarded. Introns which are directly preceded by large (longer than 30 nucleotides) insertions are

discarded as possible triplo-chimeras. The direction of an intron is determined based on the consensus sequence.

### 6.6.7 Annotating transcripts

Transcripts are annotated if at least one read connects an annotated TSS (5' end not further than 10 nucleotides away from an annotated TSS) with an annotated TES (i.e. 3' end not further than 10 nucleotides away from an annotated TES). Spliced transcripts are only annotated if there is at least one read connecting an annotated TSS with an annotated TES and contained an annotated intron.

### 6.6.8 Output files

Transcript and transcript feature annotations are exported as .gff3 files adhering to the General Feature File format version 3. Other statistics that are used for filtering are also exported as tab separated value (.tsv) tables. The outputs of multiple samples can be summed with a script that is built into the toolkit.

## 6.7 The analysis of template switching artefacts

In order to prove the effects of template switching on the analysis of transcriptional end sites and also to demonstrate the ability of the LoRTIA toolkit of efficiently filtering such artefacts, the genomic sites with potential template switching artefacts as 3' adapters (polyadenine tails) were examined. The same criteria were set for artefacts as for TES (see 5.6.5), except that here the reads tagged as "potential template switching" were used. Genomic positions that were not internal polyadenylation sites, were accepted as real TES and so were the positions where the number of not artefactual read endings in a window of ±10 nucleotides was higher than the number of artefactual read endings; all other genomic positions with at least two potentially artefactual read endings were considered artefactual TES. The common polyadenylation signals, which were described upstream of human TES (Beaudoing et al., 2000), were searched for up to 50 nucleotides upstream of both *bona fide* real and artefactual TES.

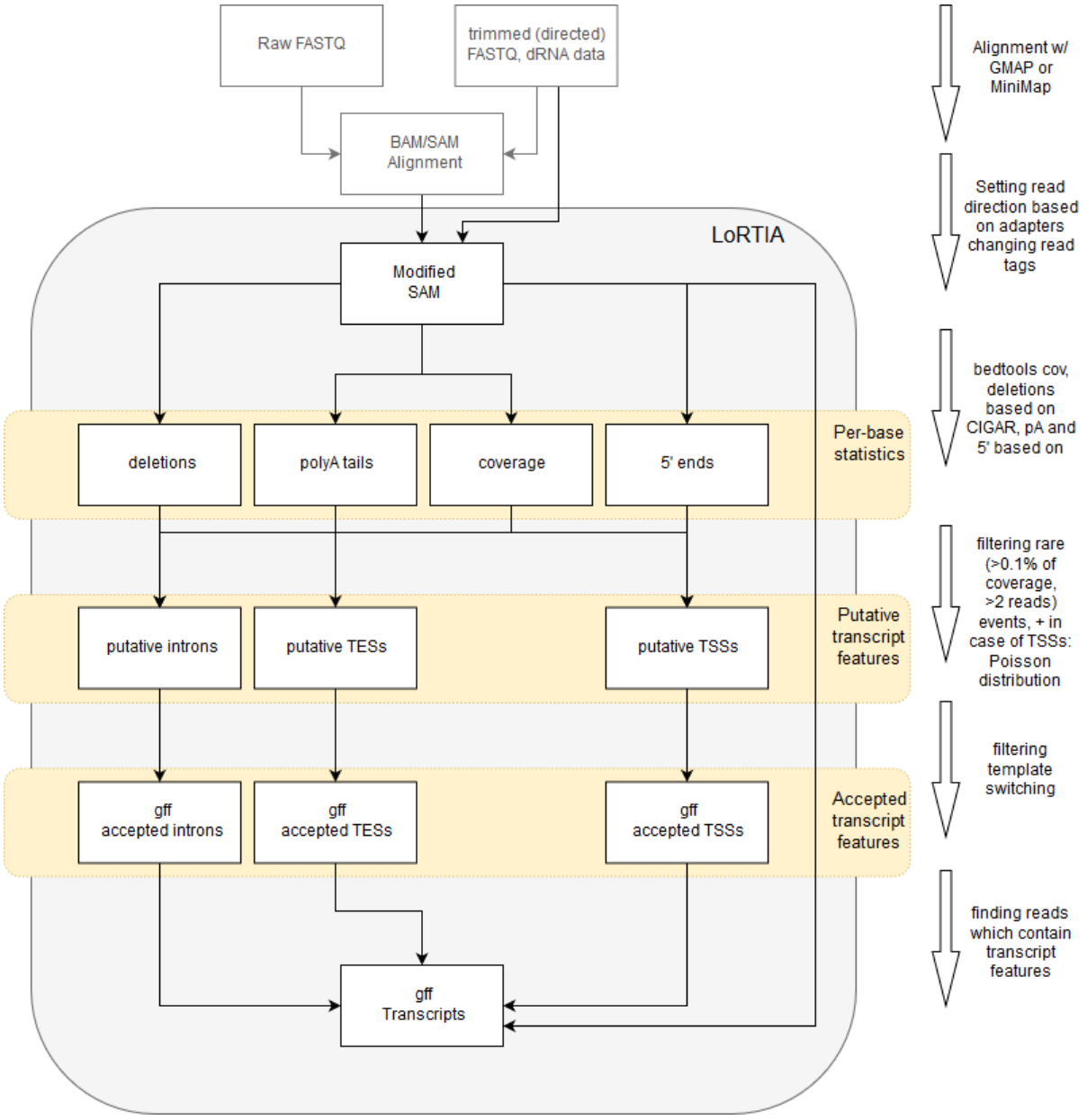


***Figure 4 The summary of the LoRTIA pipeline.*** *The software accepts aligned .bam or .sam files, and additional information as well from dRNA or preprocessed data. The software first processes the alignment, determines read orientation and set read tags that are necessary for the downstream analysis. It calculates read statistics to annotate transcript features and finally assemble the transcript isoforms. The output annotations are in .gff format. The rounded orange rectangles represent modules of the software each of which can be run/re-run on its own, making it flexible enough for handling a variety of input datasets.*

### 6.8 Visualization tools

The Integrative Genomics Viewer (IGV) (Robinson et al., 2011) was used to visualize sequencing reads. Plots were rendered using the ggplot2 (Wickham, 2016) and Gviz (Hahne and Ivanek, 2016) R packages. Nucleotide logo figures were generated using Weblogo 3 (Crooks et al., 2004). Genome annotations have been transferred and visualized with the Geneious software suite (Kearse et al., 2012).

# 7 Results

## 7.1 Mapping statistics

Our sequencing yielded 88,424 PacBio and 1,214,423 nanopore reads mapping to the HCMV genome (**Table 1**). The nanopore reads were generally shorter than PacBio reads and contained more errors (**Figure 5**). Indel errors were more common than mismatch errors.

***Table 1 The statistics of the reads mapped to the HCMV genome (LT907985)***

| Technology | Sample | Read count | Length |
|---|---|---|---|
| ONT | cap-selected and polyA-selected cDNA | 581,320 | 869.9 |
| | cap-selected random cDNA | 52,380 | 541.1 |
| | not cap-selected polyA-selected cDNA | 357,025 | 1384.9 |
| | not cap-selected polyA-selected random cDNA | 187,503 | 673.7 |
| | dRNA | 36,195 | 659.6 |
| PacBio | RSII-polyA | 45,672 | 1173.2 |
| | RSII-random | 2,509 | 923.0 |
| | Sequel | 40,243 | 1894.9 |

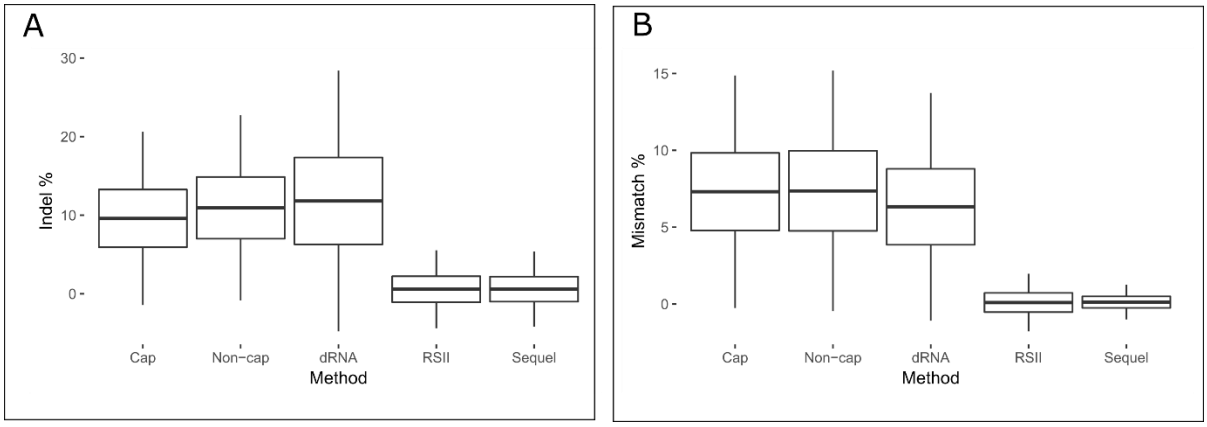


***Figure 5 Boxplots showing the different error profiles of the sequencing platforms and library preparation methods used in our studies.***

The read length distribution shows that MinION reads contained more degraded sequences than did the PacBio reads (**Figure 6**). It is also visible that size selection removed many of these degraded products. The Sequel platform produced longer reads than any other platform. The library preparation protocols for PacBio sequencing in general tend to lose a large fraction of the reads that are shorter than 1 kb. SMRT cell loading on the RSII platform

overrepresented the fraction shorter than 2kb, that is the reason size selection was recommended on the RSII platform. This bias is minimal on the Sequel platform, and it is confirmed by our data as well.

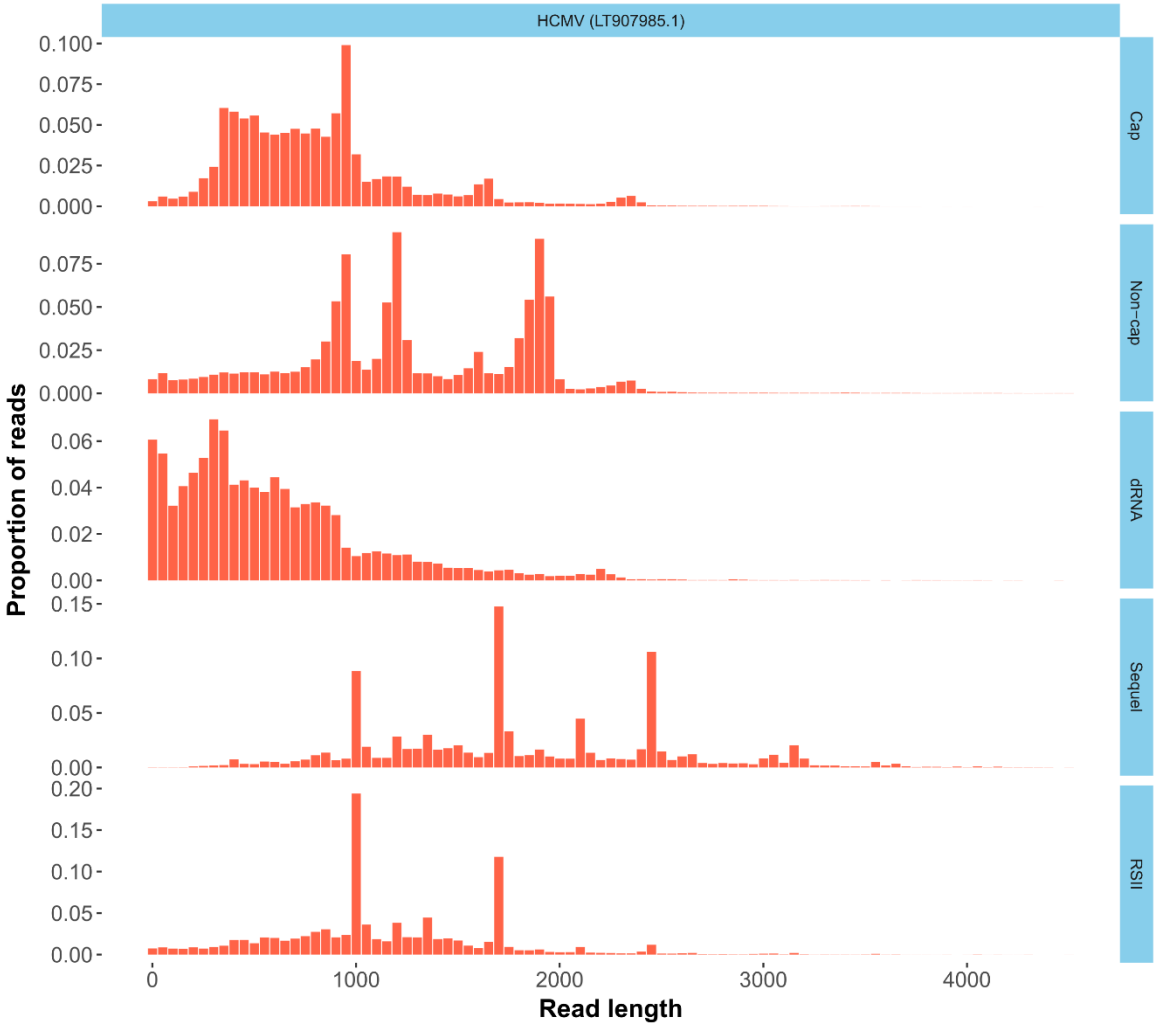


***Figure 6 Read length distribution of the poly(A)-selected sequencing runs.*** *The proportion of reads is represented for read length bins of 50 nucleotides. The high peaks in the PacBio results can be attributed to the most abundant non-coding RNAs.*

The coverage patterns of the sequencing runs were similar; the non-coding RNA molecules RNA2.7 and RNA1.2 in the long repeat (RL) region were the most abundant (**Figure 7**). These results are consistent with the findings of other groups. The coverage patterns also suggest that the (d)RNA and the cap-selected MinION reads contained more degraded sequences as the coverages of these samples are more rounded, while the other samples tended to contain more plateaus.

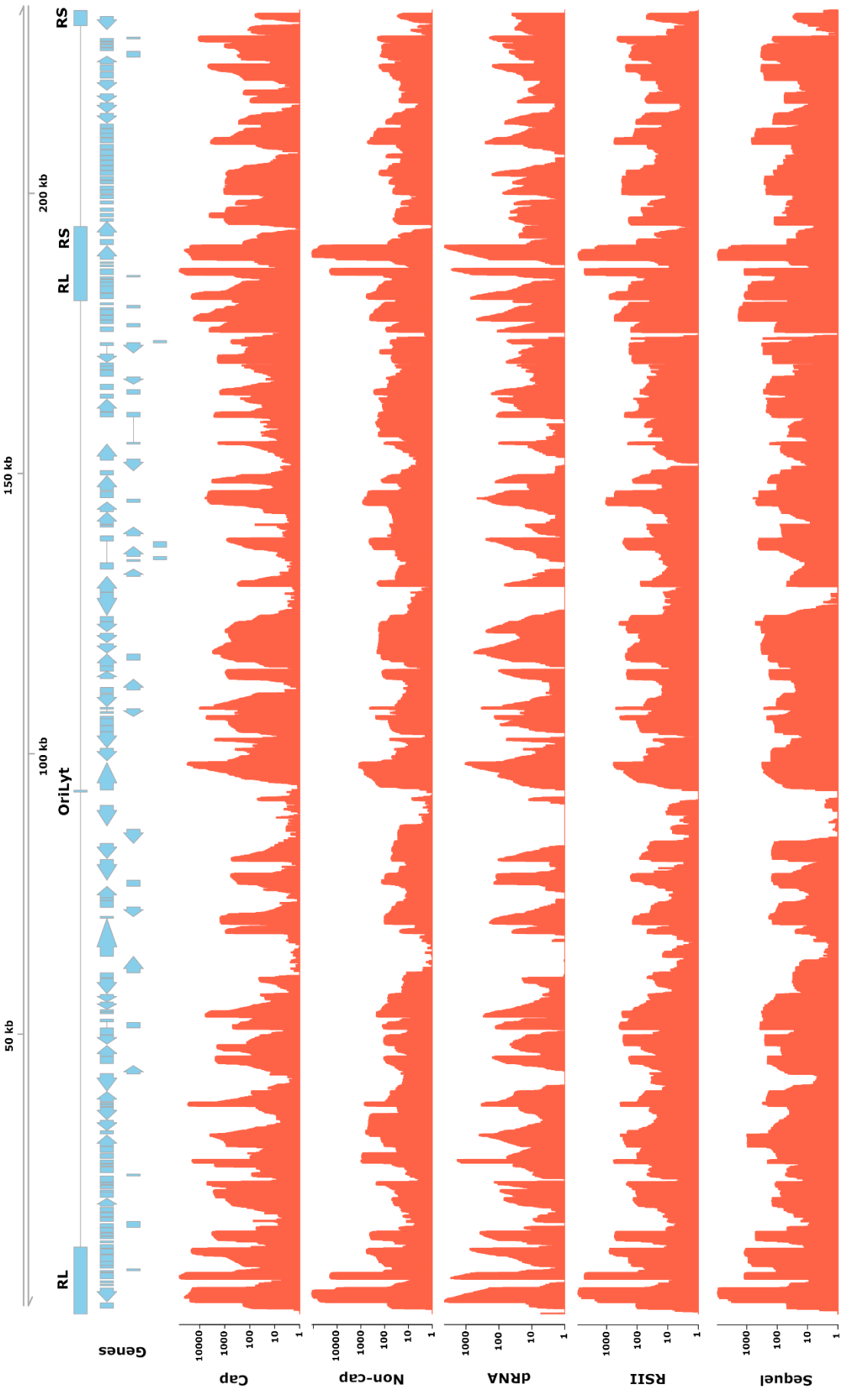
50 kb
100 kb
150 kb
200 kb
RL
OriLyt
RL
RS
RS
Genes
Cap
Non-cap
dRNA
RSII
Sequel
10000
1000
100
10
1

***Figure 7 Genome-wide coverage in the polyadenine-selected samples.*** *The larger repeat regions such as the long (RL) and short (RS) repeat sequences as well as the lytic replication of origin (OriLyt) are depicted by blue boxes. The coverage values are shown on a logarithmic scale.*

## 7.2 The identification of the virus isolate

The examined virus was obtained through the ATCC; however, it has been passaged in the Department of Medical Microbiology and Immunobiology prior to our experiments. In order to ascertain the strain of the virus, a BLAST (Altschul et al., 1990) search was conducted, where the sequencing reads were aligned against all the complete human betaherpesvirus 5 genomes in the NCBI database (until May 2017). The reads aligned to the FJ616285.1 genome showed the fewest mismatches [Table 5 in (Balázs et al., 2017b)], therefore this genome build was used as a reference for read alignments. When the genome sequence of the ATCC isolate of strain Towne HCMV was determined, it was described that the ATCC VR-977 stock of strain Towne HCMV contains two variants (Dolan et al., 2004). The longer variant (varL) was deposited in the NCBI database under the accession FJ616285.1. However, our alignments showed a larger deletion at the end of the UL region that was characteristic of varS. PCR experiments verified the existence of the deletion (**Figure 8**). The conditions of this PCR have described in details in our previous publication (Balázs et al., 2017a). Since none of the sequencing reads contained the terminal parts of the UL region, we concluded that the virus isolate used in our experiments, contained only the short variant. As only the varL genomic sequence was available in the NCBI database, we have deposited the genome sequence of varS under the accession number (LT907985) to be able to refer to genomic coordinates when annotating the viral transcriptome.

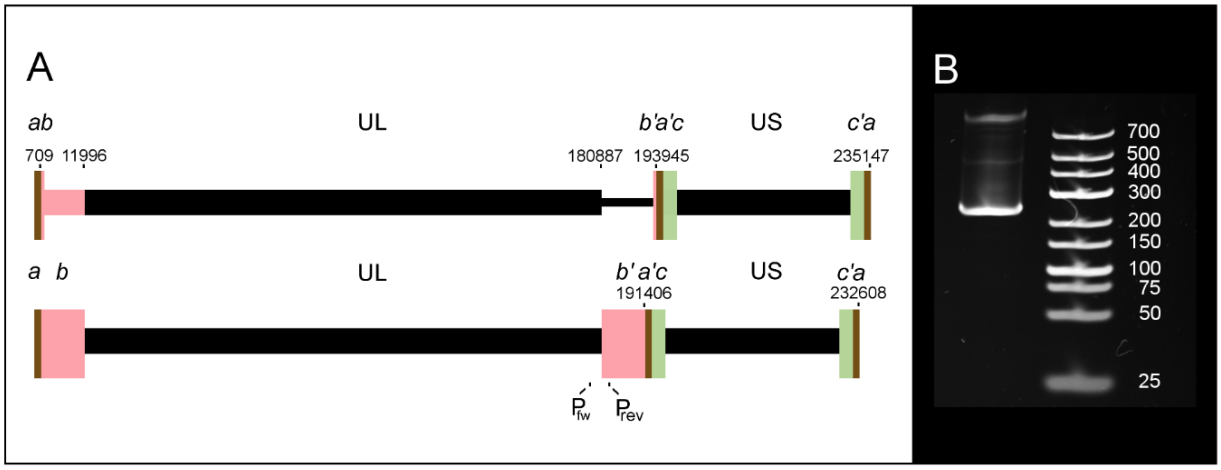


***Figure 8 The reference genome.*** *Panel A is a schematic depiction of the genomic arrangement of the virus isolates in the ATCC stock. In varS (below) the UL end region is substituted with a sequence from the beginning of the UL region which becomes the extension of the b repeat region (pink). $P_{fw}$ and $P_{rev}$ denote the PCR primers which were designed to detect the rearrangement. Panel B shows the result of the PCR.*

### 7.3 The annotation of transcriptional start sites

TSSs were identified based on the presence of the 5' adapter. The dRNA reads could not be used for TSS analysis, because 5' adapters were not ligated to the ends of the RNA molecules during dRNA sequencing. The sequencing of Nanopore reads proceeds with the help of a motor protein that ratchets the nucleic acid strands through the pore. When this motor protein reaches the end of the strand, it releases the nucleic acid, which then falls through the pore and generates nonspecific signal, therefore the very ends of a native RNA strand cannot be sequenced by dRNA sequencing (based on communication with the supplier). A similar difficulty was observed when analyzing the not cap-selected polyadenylated nanopore reads. Only 1.78% of the not cap-selected poly(A)+ and 1.69% of the not cap-selected random cDNA reads contained a recognizable adapter, in contrast to the 43.8% and 45.3% 5'-adapter-positivity of the cap-selected poly(A)+ and the cap-selected random libraries. The majority of the PacBio reads (93.8-95.5%) contained a recognizable 5' adapter. Altogether 871 transcriptional start sites were identified by our sequencing experiments, 233 of which were confirmed by at least two separate experiments. The low cross-validation rate of these sites can be partially due to the fact that the experiment which provided more than half (64%) of the

reads which contained a 5’ adapter, the cap-selected poly(A)+ cDNA library, was prepared using a biosample different from the other experiments.

### 7.4 The annotation of splice junctions

78,952 unique intron-like deletions were found in the dataset. As many of these could be indel errors, results of template switching or ligation, these deletions were filtered. Out of the 254 identified introns, 187 were validated by at least two of our sequencing experiments and another 31 were already annotated by previous short read sequencing studies (Gatherer et al., 2011; Stern-Ginossar et al., 2012). All but one of the introns contained the most common consensus splice sequence (GT/AG), and only one intron was identified which contained the second most common (GC/AG) sequence. These results are similar to the splice site usage in the host, as human splice sites tend to use the GT/AG sequence with a frequency of 98% and GC/AG with a frequency of less than 1% (Sheth et al., 2006). 105 of the accepted introns were confirmed by dRNA sequencing and none of the deletions that were filtered out by the pipeline were. The majority of the deletions detected in the reads were only detected in one read, therefore we presume that they belong to the many deletion errors that long-read platforms make. There were some deletions which were present in multiple reads, often also in multiple cDNA sequencing runs and did not contain consensus splice sequences. These deletions were flanked by short homologous sequences that can facilitate template switching (Balázs et al., 2017a), and were not detected in the dRNA sequencing experiment (**Figure 9**). As the dRNA sequencing results were not used by the pipeline to make filtering decisions, the confirmation of multiple accepted intronic sites and the lack of confirmation for the numerous sites which were filtered out, is a validation of the algorithm.


***Figure 9 Template switching artefacts may resemble introns.*** *The IGV screenshot shows a putative template switching artefact which could be detected in multiple reads in multiple cDNA sequencing runs, but not with dRNA sequencing. The blue rectangles represent aligned reads (in the reverse direction), grey lines are intron-like deletions connecting alignments. The reads which contain same false intron are highlighted with different color. The bottom line contains the genomic sequence in the region. The short homologous sequences which have probably nucleated template switching are highlighted.*

## 7.5 The annotation of transcriptional end sites

The annotation of TESs was based on the detection of polyadenine tails. The majority of the reads from polyadenine-selected cDNA libraries actually contained reads which ended in a poly(A) tail. Unfortunately, the basecalling of poly(A) tails in dRNA sequencing is hindered by the close proximity of the DNA adapters, therefore the poly(A) tail was only identified in a small proportion of dRNA reads. From the 169 identified TESs 123 were detected in at least two samples.

### 7.6 Template switching artefacts hinder the analysis of transcriptional end sites.

In order to validate the annotated TESs and to demonstrate the efficiency of the pipeline to filter out internal priming and template switching artefacts, we characterized the polyadenylation signal (PAS) usage of the *bona fide* genuine TESs and the putative template switching artefacts. The sequences surrounding the identified HCMV TESs contained the same motifs that are commonly detected around human TESs (**Figure 10**, Panel A). Upstream motifs such as the AATAAA consensus polyadenylation signal and its variant were detected 15 to 35 nucleotides upstream of the polyadenylation site. TESs usually ended with an adenine, and were followed by a GU-rich region, which is a known signal of the cleavage factor (Pérez Cañadillas et al., 2003). The putative artefactual polyadenylation sites, were identified by reads that contained polyadenylation tails, but were immediately preceded by A-rich regions, which may had triggered template switching or internal priming. These putative artefacts had neither the upstream nor the downstream sequence motifs that were commonly observed in TESs. The PAS usage of genuine HCMV TESs resembled that of the human TESs' (**Figure 10**, Panel B), however the artefactual sites were rarely preceded by consensus PASs, and when they were, these PASs were less likely to fall into the expected -15 to -35 nucleotide range than those of the genuine ones (**Figure 10**, Panel C).

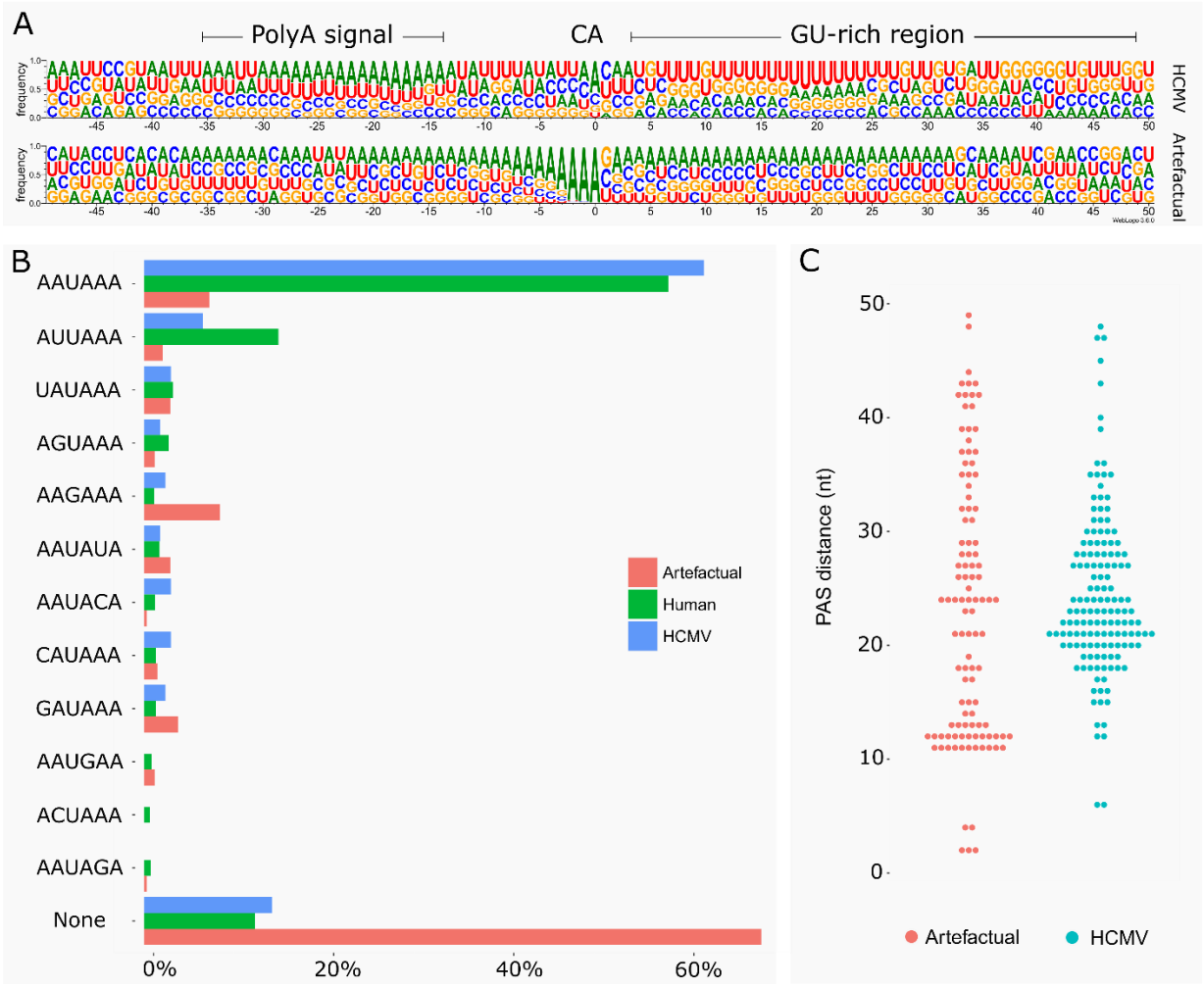


***Figure 10 PAS usage of putative artefactual and transcriptional end sites.*** *Panel A shows a Weblogo* (Crooks et al., 2004) *of the ±50 nucleotide vicinity of the identified HCMV TESs and putative artefactual TESs. The polyadenylation site is at position 0. Panel B shows the frequency of each PAS. The human PAS usage data is based on* (Beaudoing et al., 2000). *Panel C shows the distance of the identified PASs from the TESs.*

Direct RNA sequencing is devoid of reverse-transcription and PCR artefacts. Therefore, to further test the validity of our results, we tested whether the spurious TESs can be differentiated by dRNA sequencing. The basecalling of the poly(A) tails was often unsuccessful (as described above), consequently even the sites that were identified by the pipeline as real TESs, were rarely confirmed by polyadenylated dRNA reads. However, as shown in **Figure 11**, the dRNA reads very often contained the complete 3' sequence of the transcripts, only the basecalling of the homopolymer (A)s (i.e. the polyadenine tail) was

disrupted. Based on this information, the 3’ ends of dRNA reads can be used to validate TESs, even if the reads do not contain a poly(A) tail.

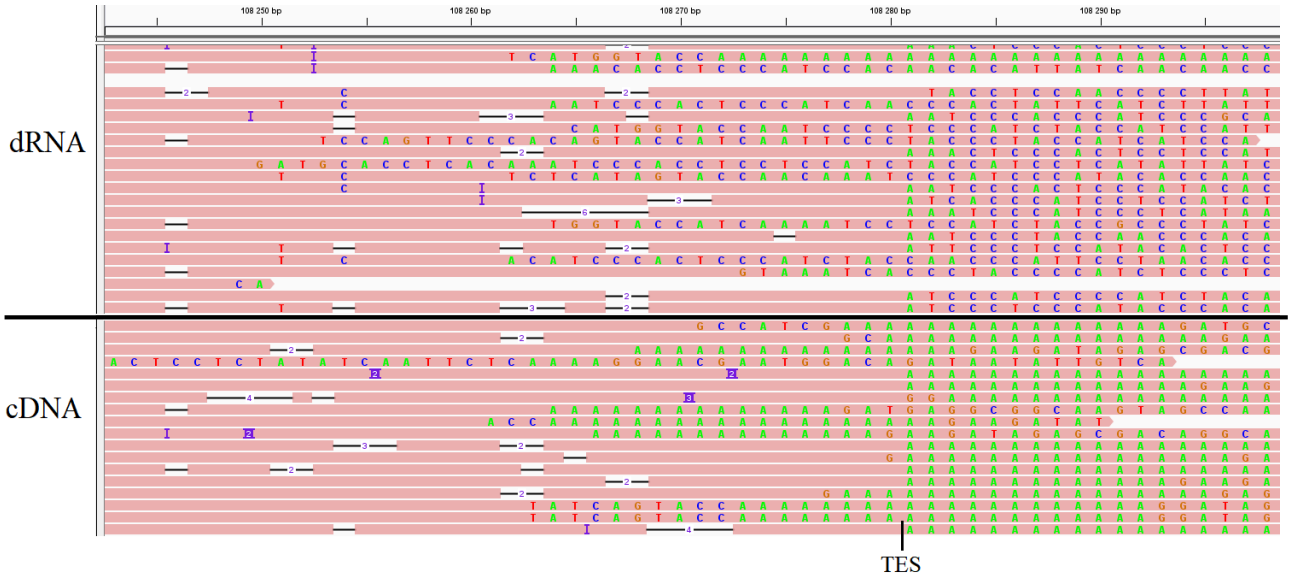


***Figure 11 The basecalling of the poly(A) tails is impaired in dRNA sequencing.*** *The IGV screenshot shows dRNA (above) and cDNA (below) reads from the same genomic region. The pink lines represent the aligned parts of the reads. The colorful letters represent unmapped bases. A long sequence of unmapped (A)s is visible in the cDNA sequencing dataset, but not in the dRNA dataset. However, the aligned parts of the dRNA reads also end at the TES, and a nonspecific sequence follows.*

The putative artefactual polyadenylation sites were less frequently utilized than the *bona fide* real sites. However, a strong linear correlation was observed between the logarithm of the proportion of reads ending at a given artefactual position and the number consecutive adenines preceding the site (**Figure 12**). The heatmap representation of read endings in the neighboring locations around *bona fide* real HCMV and artefactual TESs in **Figure 13** shows that most artefactual positions were detected in multiple cDNA sequencing samples similar to the accepted TESs. The accepted TESs were often confirmed by polyadenylated dRNA reads or at least by specific signals of read endings in the dRNA sequencing runs, however, the putative artefactual cleavage sites were not. Another characteristic of the real TESs is that the cleavage site is distributive (i.e. the transcript is not always cleaved at the exact same nucleotide, but usually in the close proximity of the most frequent site) (Sheppard et al., 2013). The artefacts, on the other hand, were usually detected at the exact same position, which is due to the fact that these are derived from (A)-rich regions (the criterion for identifying potential

template switching artefacts was a minimum of 3 (A)s) (Balázs et al., 2017a). If the majority of these artefacts were caused by template switching, the artefactual polyadenine tails should be shorter than the polyadenine tails at the real TESs, because the oligod(T)-primers used in our experiments were 20 nucleotides long and real polyadenine tails are usually longer than that. However, we found that the length distribution of artefactual and real poly(A) tails were very similar in all of our cDNA sequencing samples (**Figure 14**).

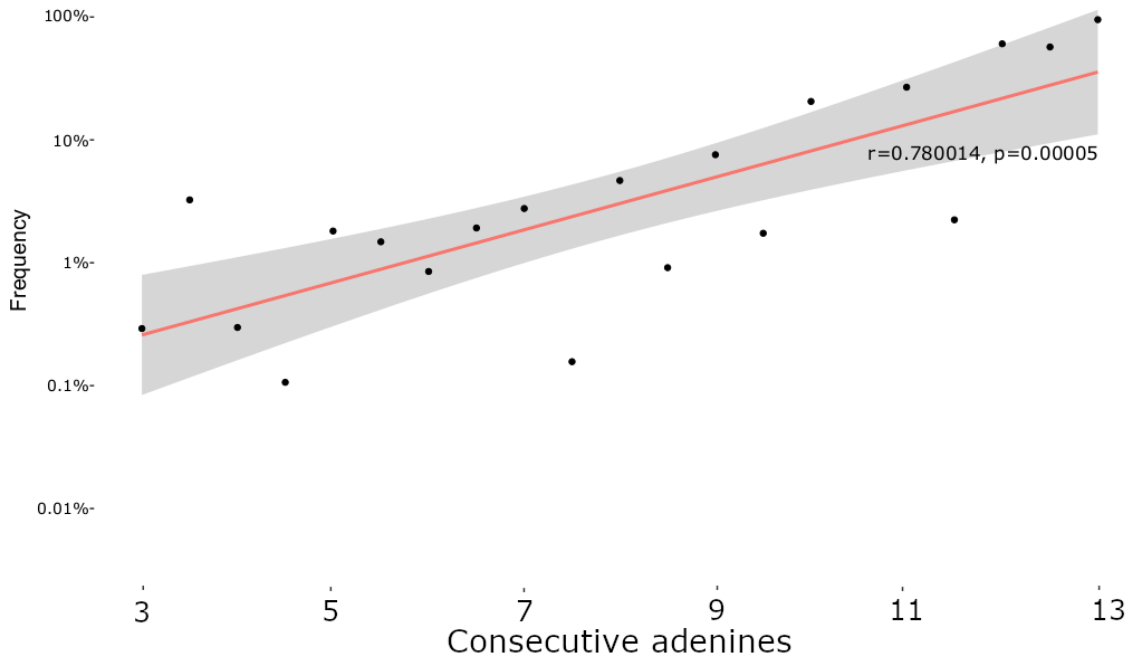


***Figure 12 The number of consecutive adenines correlates with the proportion of artefactually polyadenylated reads.*** *The frequency of polyadenylation is shown on a logarithmic scale. The number of consecutive adenines immediately upstream of a position was determined based on the genomic sequence. When a homopolymer stretch of adenines was interrupted by other nucleotides, the other nucleotides were counted as -1.5; for instance, the sequence AAAGAA would be counted as 3.5 consecutive adenines. A red trendline with the Pearson correlation coefficient is shown.*

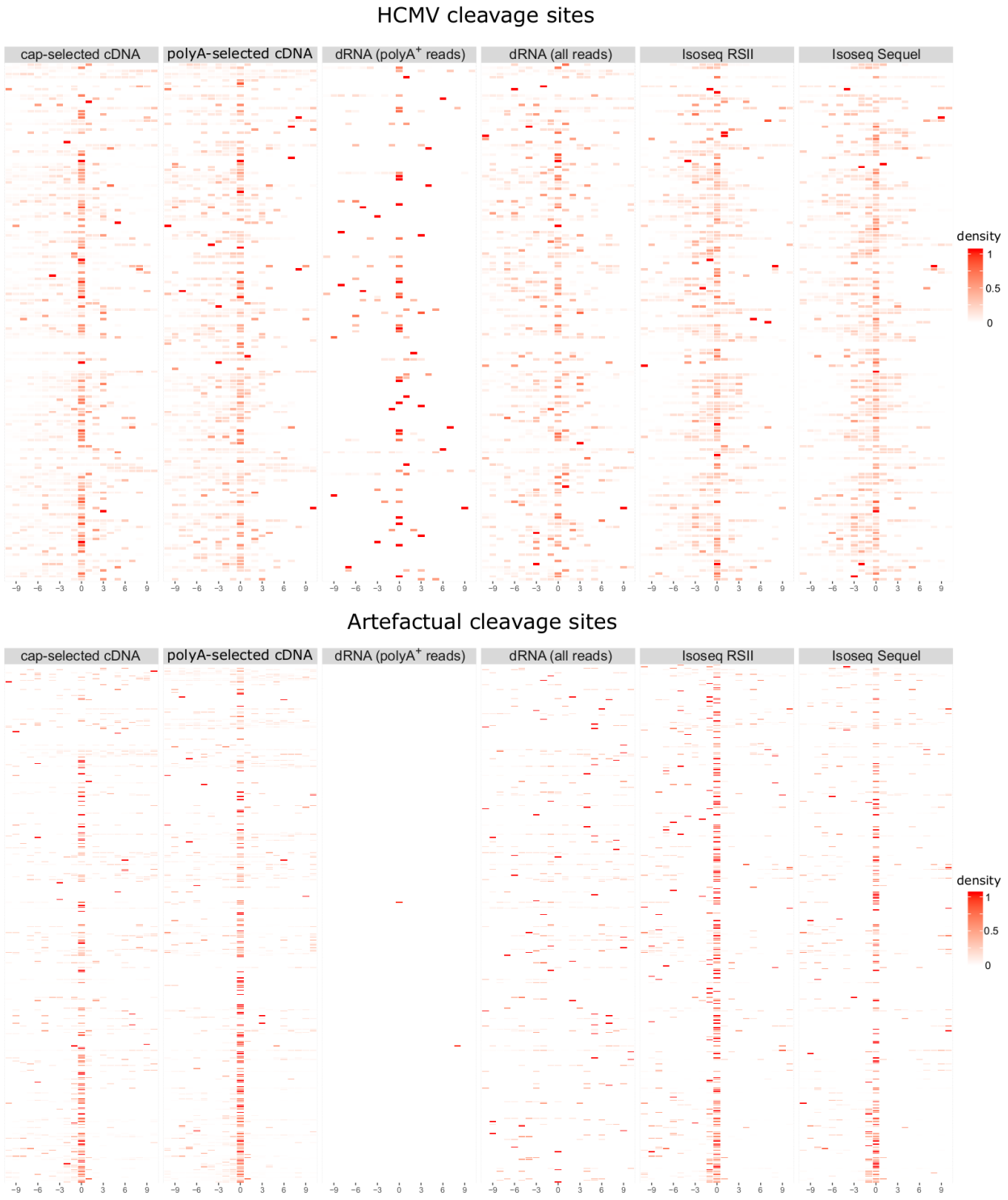


***Figure 13 Artefactual TESs can be differentiated from real ones, using dRNA sequencing.*** *Heatmap of the 3' ends of reads in the vicinity of the annotated TESs (above) and putative artefactual polyadenylation sites (below). Each row is a TES, or a putative artefact and columns are the surrounding nucleotides depicted for each poly(A)-selected sequencing*

*experiment. For all cDNA experiments, only the reads which ended with a poly(A) tail were calculated. The dRNA sequencing results are presented in two ribbons, one only showing the reads that ended with an identifiable poly(A) tail (specific, but only few reads), and one that shows all the 3' ends (less specific, but more reads). The density value for each position is calculated by dividing the number of reads which ended at a given position by the sum of the reads which ended in the vicinity of the TES.*

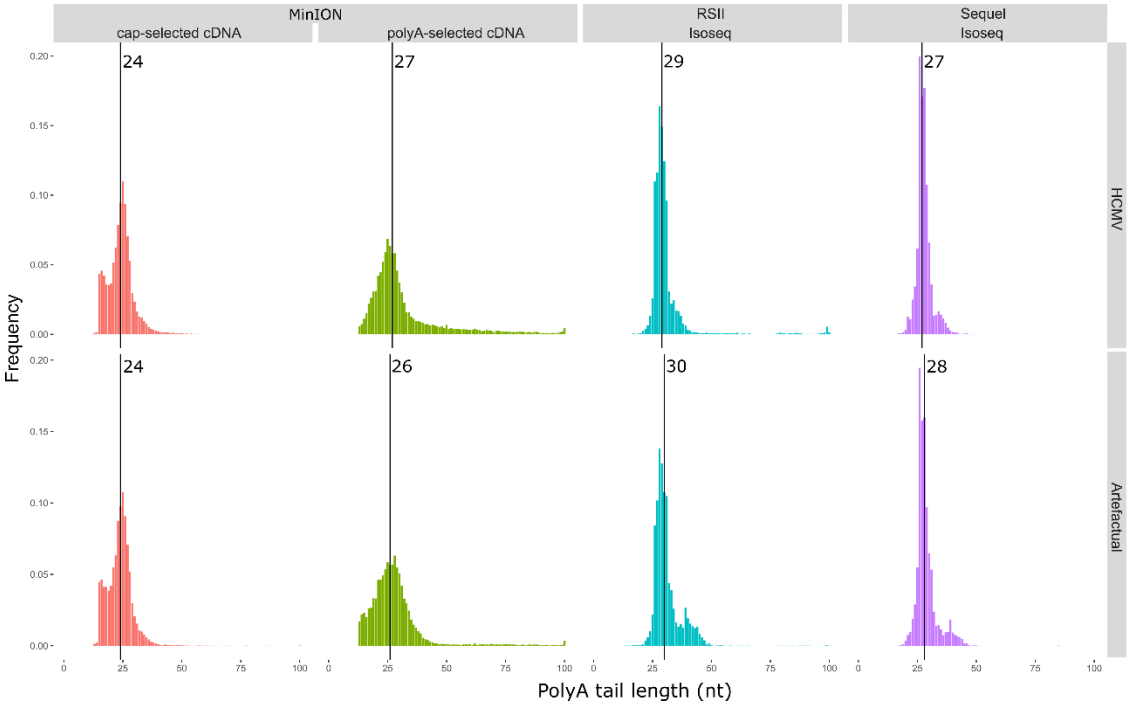


***Figure 14 Poly(A) tail length distributions in the poly(A)-selected cDNA samples.*** *The length of the 3' terminal (A) mismatches were measured at annotated TESs and at putative artefactual sites as well. Black vertical lines depict the median values.*

## 7.7 The annotation of transcript isoforms

Based on the annotated transcript features which were confirmed by at least two sequencing experiments, 440 transcript isoforms were annotated. 377 of these isoforms were novel isoforms. 104 already described transcripts were confirmed by our annotations, while 25 were not. The majority of the identified transcripts were confirmed by at least two sequencing runs (**Figure 15**). Even though the PacBio platforms had a substantially lower throughput than the nanopore platform, the PacBio platforms identified the most transcript isoforms.

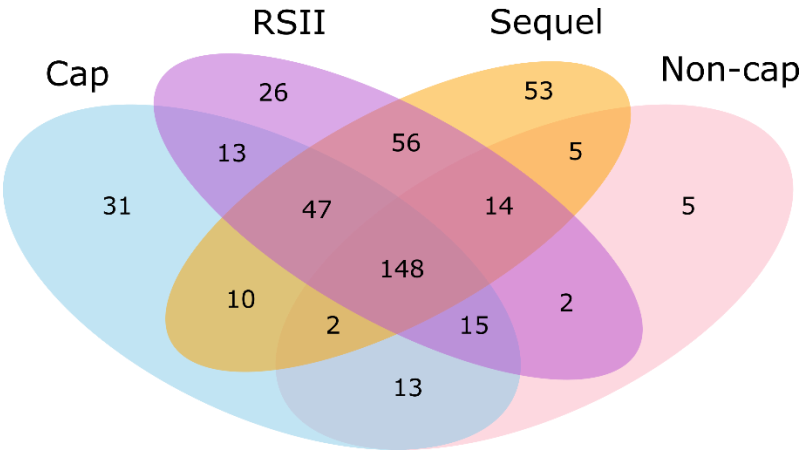


***Figure 15 Venn diagram of the number of transcripts identified by each sequencing method.*** *Only those sequencing runs are depicted, which provided full-length transcript information.*

The majority of the isoforms were polycistronic isoforms, 5'-UTR or splice isoforms. Alternative polyadenylation was also present, however it produced much less variation. The different sequencing methods identified similar proportions of the types of transcript isoforms (**Figure 16**).

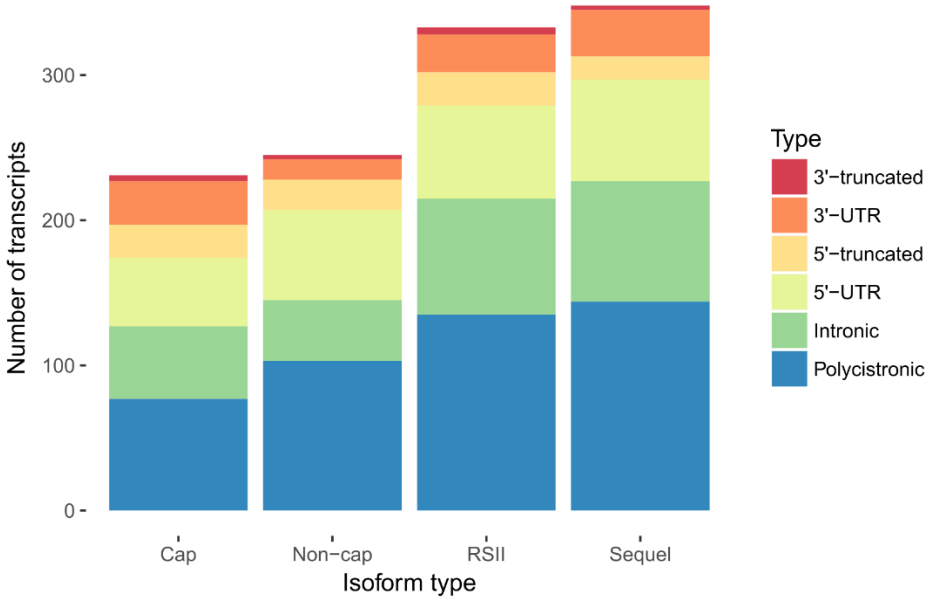


***Figure 16 The types of transcript isoforms.*** *The polycistronic isoforms differed in the number of coding sequences they were carrying, the intronic isoforms used different splice junctions. The UTR-isoforms differed in the length of the UTRs, while the truncated isoforms contained a shorter version of the main protein.*

## 7.8 Novel HCMV transcripts

Most of the transcripts detected in our experiments were isoforms of already known transcripts or previously uncharacterized transcripts of known genes (**Figure 17**). However, we have also described novel transcripts, antisense to eight known genes: UL20, UL36, UL38, UL54, UL115, US1, US17 and US30 (Balázs et al., 2017a). Another novel intergenic transcript, named RS2, was detected partially antisense to RS1 in the short repeat region.

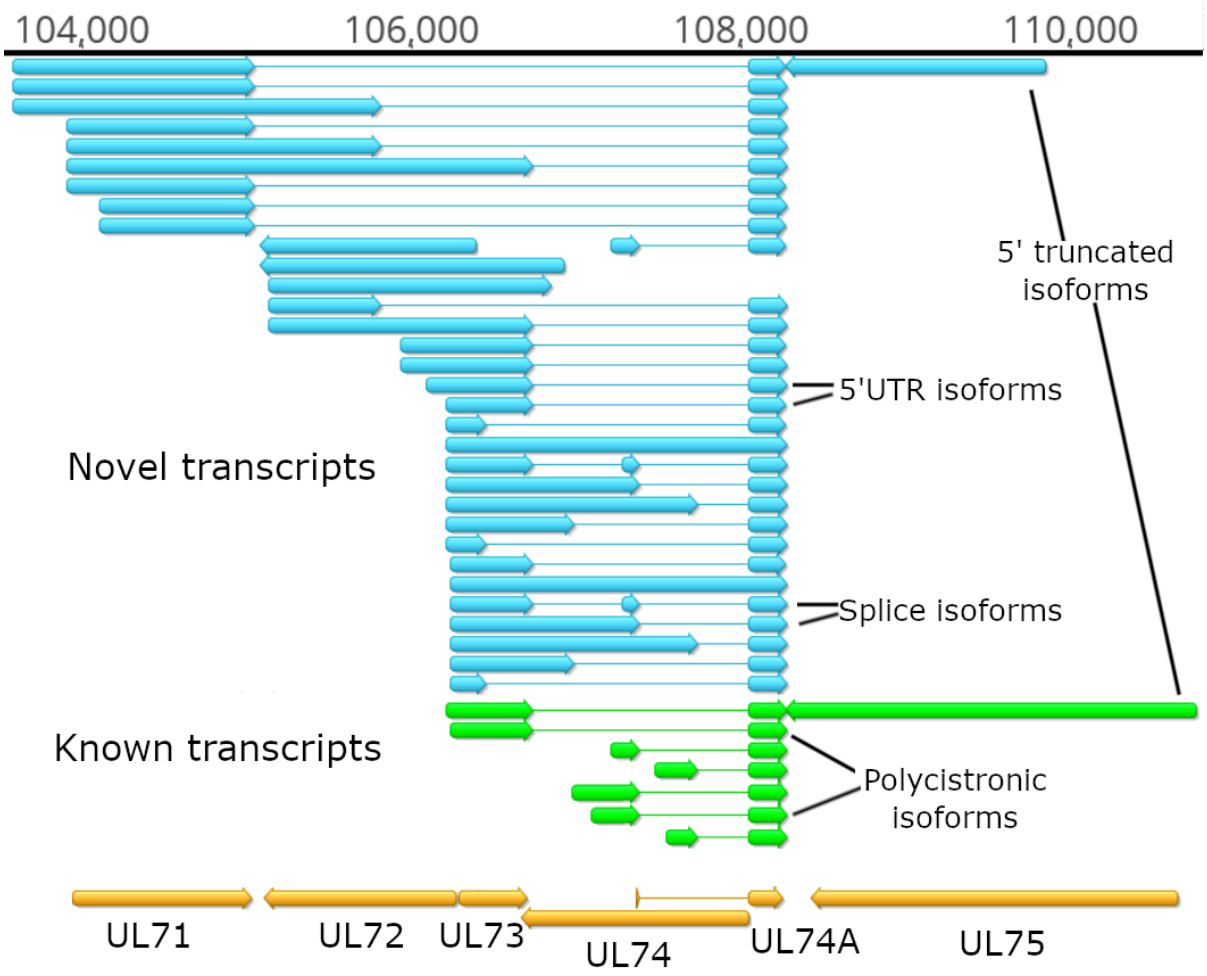


***Figure 17 Transcript isoforms of the HCMV UL71-UL75 region.*** *The genomic positions are marked on the top part of the figure. The novel transcript isoforms identified by our experiments are depicted with blue, already known and detected transcript isoforms are shown with green, while the protein coding sequences are yellow. Examples for different types of isoforms are labelled on the right. Thin lines represent introns.*

# 8 Discussion

We have sequenced RNA and cDNA libraries prepared from HCMV-infected cells to characterize the lytic HCMV transcriptome. Various library preparation methods were used in order to capture multiple aspects of the transcripts. The majority of the reads stem from poly(A)-selected libraries, and therefore our studies mainly focused on the polyadenylated fraction of RNAs. The data we have generated is one of the few openly accessible long-read RNA sequencing datasets. Raw and mapped data are available on the website of the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under the accession numbers PRJEB22072 and PRJEB25680. The transcriptome annotation is available under the accession number LT907985. The dataset can be used to improve bioinformatic tools for the analysis of long-read sequencing data or to better understand the molecular biology of HCMV.

We developed a software toolkit for the analysis of long-read RNA sequencing data. The toolkit is compatible with data from the two main long-read sequencing technologies. The program identifies transcript features such as TSSs, TESs and introns and assembles the transcripts based on these features. The toolkit efficiently handles noisy long-read sequencing data and does not require any previous genome annotation, although it does require a genome sequence. The pipeline is flexible, the many filtering parameters can be changed by the user, which makes it adaptable to different sequencing platforms and different experimental settings.

The software also filters template switching artefacts. We have validated the efficacy of the pipeline by comparing the putative artefactual TESs to the accepted TESs. Artefactual TESs did not contain the consensus regulatory elements of polyadenylation, which could clearly be observed at the *bona fide* real TESs. This line of evidence raises suspicion towards the TESs which were found in (A)-rich regions, however, it would also be possible that such (A)-rich regions represent a different, weaker kind of signal for alternative polyadenylation. In order to ascertain the artefactual nature of these TESs, we have compared the dRNA sequencing results to the cDNA sequencing results. Even though several of the putative artefactual sites were detected in multiple cDNA sequencing experiments, they were not detected in the dRNA experiments. Since dRNA experiments are unaffected by reverse transcription and PCR artefacts, we can argue that the putative artefactual sites that the pipeline filters out, are indeed technical artefacts. The LoRTIA pipeline uses a much stricter filtering

method than previous RNA sequencing or expressed sequence tag analysis methods would use. Such methods assume that internal priming is the main cause of polyadenylation artefacts and generally discard genomic locations with six or more consecutive adenines. However, our method assumes that template switching is also an important factor in the generation of such artefacts and therefore labels genomic locations which contain three or more consecutive (A)s as potential template switching artefacts and filters such locations based on the distribution read endings compared to coverage and whether reads contained polyadenine tails in the vicinity of the site, at locations which contain less adenines. The fact that the length of the poly(A) tails was not shorter at the artefactual sites supports the assumption that many of the artefacts arose from template switching. As 20-nucleotide long oligod(T) primers were used for reverse transcription, if the artefacts were generated by internal priming, artefactual poly(A) tails were supposed to be shorter. Numerous artefacts were detected at (A)-rich regions composed of as few as three consecutive (A)s, which is not likely to have been caused by internal priming. We regard these findings as a validation of an algorithm that assumes template switching is an important factor in the generation of such artefacts.

There can be numerous explanations as to why template switching artefacts were not considered to be an important source of artefacts by previous studies. One reason might be that template switching has been shown to occur more frequently during reverse transcription at lower temperatures (Mader et al., 2001) and the SMART method (Zhu et al., 2001), commonly used for the preparation of long-read sequencing libraries, requires 42°C for the synthesis of the second cDNA strand. Short-read RNA sequencing libraries but especially poly(A) Seq libraries do not require such conditions, and their reverse transcription can be carried out at higher temperatures. Another reason might be that short-read sequencing studies usually produce a large number of reads; therefore, high threshold can be set in order to filter out technical artefacts. Long-read sequencing on the other hand generally has a low throughput and the analysis has to rely on less reads. For example, **Figure 12** showed that artefactual sites that contained only six or less consecutive adenines, usually only lead to template switching in fewer than 1% of the reads. Accordingly, a short-read sequencing study with a high coverage could efficiently filter out the majority of such positions.

Out pipeline has identified most of the previously annotated transcripts as well as a large number of novel transcripts. There were several already described transcripts or transcript features that our experiments did not detect. The reason for this may have been to the relatively low throughput of our sequencing or that some transcript isoforms are only expressed in the immediate-early or early phases of infection and our dataset presumably only contained a low number of reads from these time points. Another reason may be that some isoforms are specific to some strains and are not expressed in the strain that we used. Strain Towne is a highly passaged strain with numerous mutations, compared to clinical HCMV strains (Dolan et al., 2004). We decided to use the Towne strain as it was the strain with the highest number of already annotated transcripts.

Nevertheless, our experiments revealed a high transcript diversity in HCMV. A similar amount of diversity has been revealed by long-read sequencing in other viruses and higher order eukaryotes as well (Abdel-Ghany et al., 2016; Moldován et al., 2018; O'Grady et al., 2016; Sharon et al., 2013; Tombácz et al., 2016, 2017b; Wang et al., 2016; Workman et al., 2018). The functions of the novel transcripts are yet unknown. Some of the transcripts are antisense to other genes, while the majority of the transcripts contain large already annotated protein coding sequences, therefore they probably code for proteins. The reason why HCMV and the other organisms produce so many different transcript variants from the same gene though is still to be explained.

One explanation could be that different isoforms are expressed with different kinetics, such as the transcripts of the ICP36, the major DNA-binding protein (Isomura et al., 2008). Long-read sequencing can help in answering this question, because long-read sequencing can differentiate isoforms more efficiently, therefore quantitative long-read sequencing could facilitate the functional annotation of transcript isoforms (Tombácz et al., 2017a). We have detected numerous antisense and also partially antisense or overlapping transcripts; it is possible that some isoforms regulate the neighboring genes through transcriptional interference (Eszterhas et al., 2002). The transcriptional interference network hypothesis says that the overlapping transcripts form genetic modules and the expression of each gene influences the expression of the other genes in that module (Boldogköi, 2012). Extensive antisense transcription from a large part of the HCMV genome had already been described using short-

read sequencing (Gatherer et al., 2011), but this transcriptional activity has not been attributed to distinct transcripts. Ribosome profiling studies have revealed extensive translation from short upstream ORFs (Stern-Ginossar et al., 2012). We have found that many transcript isoforms differ in whether they contain upstream ORFs (uORF) upstream of the main ORF or not (Dataset S7 in (Balázs et al., 2017a)). Upstream ORFs have been shown to be able to control the translation of downstream genes in HCMV (Geballe and Mocarski, 1988) and in other organisms as well (Barbosa and Romao, 2014; Nyikó et al., 2009; Vilela and McCarthy, 2003; Wethmar, 2014). Polycistronic transcripts are very common in herpesviruses, and uORFs have been shown to be able to change which polypeptides are translated from a transcript (Kronstad et al., 2013). Another finding in the comparison of ribosome profiling and long-read sequencing data was that the 5'-truncated transcripts contain N-terminally truncated polypeptides which are translated and may have different functions from the main proteins. However, it is also possible that much of the transcript diversity is merely transcriptional noise and that many transcript isoforms do not have differential effects. Further studies are needed to elucidate the function of each of the transcript isoforms.

We also compared the performance of multiple long-read sequencing platforms in the analysis of the HCMV transcriptome. We confirmed the improvement in the throughput of the Sequel platform compared to the RSII platform and we, similarly to others (Weirather et al., 2017), have obtained substantially more reads from the nanopore platform than from the PacBio platform. Notwithstanding the higher throughput of the nanopore platform, the PacBio reads were more likely to depict full-length transcripts. Even though we have obtained much more transcripts with the nanopore technology, more transcripts were detected by the PacBio technology. This may be due to the effect of the IsoSeq library preparation that eliminates short cDNA fragments, similar effects can be reached on the ONT platform by applying size selection. However, the ability of the ONT platform to sequence short nucleic acid fragments is not necessarily a disability; many previously described short transcripts (longer than 400 nucleotides, but shorter than 1000 nucleotides) could not be detected by the PacBio platforms but were identifiable by MinION sequencing.

Native RNA sequencing on the ONT platform still has several drawbacks such as its low throughput, the impaired basecalling of the 3' end and the technology's inability to

sequence the 5’ ends of the RNA molecules. Despite all of these weaknesses, dRNA sequencing is immune to a number of artefacts that plague cDNA sequencing. By applying dRNA sequencing we managed to ascertain the artefactual nature of numerous TESs and introns.

# 9 Conclusions

We have applied long-read RNA sequencing in the study of the HCMV transcriptome. Our results have more than tripled the number of annotated HCMV transcripts. Cross-platform validation and the confirmation of dRNA sequencing gives these novel features high confidence. Using long-read RNA sequencing data we were able to draw a more detailed map of the HCMV transcriptome, which is instrumental both for the analysis of viral gene expression and for understanding the molecular mechanisms of infection. The publicly available data may facilitate the detailed analysis of HCMV genes as well as provide potential targets for disease control.

We have developed and published a bioinformatic toolkit for the analysis of long-read sequencing data, which can be used for the analysis of long-read RNA sequencing data from various platforms and from various organisms. The flexibility of the pipeline allows for the examination of simple viral or higher order eukaryotic transcriptomes as well.

# 10 Acknowledgements

I have received funding to support some of the projects presented in the thesis from the Hungarian Ministry of Human Capacities in the form of the NTP-NFTÖ-17B scholarship for the ONT cDNA sequencing experiments and the UNKP-18-3-IV-SZTE-2 scholarship for the development of the bioinformatics pipeline to analyze long-read RNA sequencing data.

I would like to thank Dr. Klára Megyeri from the Department of Microbiology and Immunobiology for the virus and the cells that she supplied.

I am also grateful for the opportunity of having been able to work with PacBio data, this would not have been possible without the support of Dr Donald Sharon and Prof Michael Snyder from the Department of Genetics at the University of Stanford.

I would like to express my gratitude to all my colleagues at the Department of Medical Biology for their constant help and support throughout the years and also during the writing of this thesis. I would like say special thanks to Marianna Ábrahám, for her assistance in the lab, to Dr Zsolt Csabai, who did most of the wet lab work for the presented studies, to Dr Dóra Tombácz, who performed the PacBio sequencing experiments and also helped with the preparations of the nanopore sequencing, to Norbert Moldován, who helped with the nanopore sequencing in the lab, assisted with the analyses and helped brainstorming ideas, to Dr Attila Szűcs, who carried out much of the analysis of the early data as well as helped outline some aspects of the pipeline and finally to Zsolt Boldogkői, who planned the studies and supervised the writing of the thesis.

# 11 References:

Aaronson, J. S., Eckman, B., Blevins, R. A., Borkowski, J. A., Myerson, J., Imran, S., et al. (1996). Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res.* 6, 829–45. doi:10.1101/GR.6.9.829.

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., et al. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7, 11706. doi:10.1038/ncomms11706.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of Drosophila melanogaster. *Science* 287, 2185–95. doi:10.1126/SCIENCE.287.5461.2185.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2.

Arend, K. C., Ziehr, B., Vincent, H. A., and Moorman, N. J. (2016). Multiple Transcripts Encode Full-Length Human Cytomegalovirus IE1 and IE2 Proteins during Lytic Infection. *J. Virol.* 90, 8855–65. doi:10.1128/JVI.00741-16.

Balázs, Z. (2018). LoRTIA. Available at: https://github.com/zsolt-balazs/LoRTIA [Accessed January 16, 2019].

Balázs, Z., Tombácz, D., Szűcs, A., Csabai, Z., Megyeri, K., Petrov, A. N., et al. (2017a). Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials. *Sci. Rep.* 7, 15989. doi:doi 10.1038/s41598-017-16262-z.

Balázs, Z., Tombácz, D., Szűcs, A., Snyder, M., and Boldogkői, Z. (2017b). Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform. *Sci. Data* 4, 170194. doi:10.1038/sdata.2017.194.

Barbosa, C., and Romao, L. (2014). Translation of the human erythropoietin transcript is regulated by an upstream open reading frame in response to hypoxia. *RNA* 20, 594–608. doi:10.1261/rna.040915.113.

Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* 10, 1001–1010.

doi:10.1101/gr.10.7.1001.

Boldogköi, Z. (2012). Transcriptional interference networks coordinate the expression of functionally related genes clustered in the same genomic loci. *Front. Genet.* 3, 122. doi:10.3389/fgene.2012.00122.

Brakenhoff, R. H., Schoenmakers, J. G. G., and Lubsen, N. H. (1991). Chimeric cDNA clones: a novel PCR artifact. *Nucleic Acids Res.* 19, 1949–1949. doi:10.1093/nar/19.8.1949.

Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., and Blake, J. A. (2007). The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.* 36, D724–D728. doi:10.1093/nar/gkm961.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* 40, 722–729. doi:10.1038/ng.128.

Cannon, M. J., Schmid, D. S., and Hyde, T. B. (2010). Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Rev. Med. Virol.* 20, 202–213. doi:10.1002/rmv.655.

Caviness, K., Cicchini, L., Rak, M., Umashankar, M., and Goodrum, F. (2014). Complex expression of the UL136 gene of human cytomegalovirus results in multiple protein isoforms with unique roles in replication. *J. Virol.* 88, 14412–25. doi:10.1128/JVI.02711-14.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–3. doi:10.1093/bioinformatics/btp163.

Cocquet, J., Chong, A., Zhang, G., and Veitia, R. A. (2006). Reverse transcriptase template switching and false alternative transcripts. doi:10.1016/j.ygeno.2005.12.013.

Cramer, P., Pesce, C. G., Baralle, F. E., and Kornblihtt, A. R. (1997). Functional association between promoter structure and transcript alternative splicing. *Proc. Natl. Acad. Sci.* 94, 11456–11460. doi:10.1073/pnas.94.21.11456.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–90. doi:10.1101/gr.849004.

Dang, Q., and Hu, W. S. (2001). Effects of homology length in the repeat region on minus-

strand DNA transfer and retroviral replication. *J. Virol.* 75, 809–20. doi:10.1128/JVI.75.2.809-820.2001.

Davis, N. L., King, C. C., and Kourtis, A. P. (2017). Cytomegalovirus infection in pregnancy. *Birth Defects Res.* 109, 336–346. doi:10.1002/bdra.23601.

Davison, A. J., Dolan, A., Akter, P., Addison, C., Dargan, D. J., Alcendor, D. J., et al. (2003). The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J. Gen. Virol.* 84, 17–28. doi:10.1099/vir.0.18606-0.

Dolan, A., Cunningham, C., Hector, R. D., Hassan-Walker, A. F., Lee, L., Addison, C., et al. (2004). Genetic content of wild-type human cytomegalovirus. *J. Gen. Virol.* 85, 1301–1312. doi:10.1099/vir.0.79888-0.

Edwalds-Gilbert, G., Veraldi, K. L., and Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.* 25, 2547–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9185563 [Accessed February 4, 2016].

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., et al. (2012). Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* 7, e47768. doi:10.1371/journal.pone.0047768.

Eszterhas, S. K., Bouhassira, E. E., Martin, D. I. K., and Fiering, S. (2002). Transcriptional Interference by Independently Regulated Genes Occurs in Any Relative Arrangement of the Genes and Is Influenced by Chromosomal Integration Position. *Mol. Cell. Biol.* 22, 469–79. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11756543 [Accessed November 27, 2017].

Fields, B. N., Knipe, D. M. (David M., and Howley, P. M. (2013). *Fields virology*. 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins.

Gao, S., Gao, S., Ruan, Q., Ruan, Q., Ma, Y., Ma, Y., et al. (2015). Validation of three splice donor and three splice acceptor sites for regulating four novel low-abundance spliced transcripts of human cytomegalovirus UL21.5 gene locus. *Int. J. Mol. Med.* 35, 253–262. Available at: https://www.spandidos-publications.com/ijmm/35/1/253 [Accessed August 15, 2017].

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–

206. doi:10.1038/nmeth.4577.

Gardini, A. (2017). “Global Run-On Sequencing (GRO-Seq),” in *Methods in molecular biology (Clifton, N.J.)*, 111–120. doi:10.1007/978-1-4939-4035-6_9.

Gatherer, D., Seirafian, S., Cunningham, C., Holton, M., Dargan, D. J., Baluchova, K., et al. (2011). High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* 108, 19755–60. doi:10.1073/pnas.1115861108.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J. M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.* 8, 524–30. doi:10.1101/GR.8.5.524.

Geballe, A. P., and Mocarski, E. S. (1988). Translational control of cytomegalovirus gene expression is mediated by upstream AUG codons. *J. Virol.* 62, 3334–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2841486 [Accessed May 14, 2017].

Geiszt, M., Lekstrom, K., and Leto, T. L. (2004). Analysis of mRNA Transcripts from the NAD(P)H Oxidase 1 ( *Nox1* ) Gene. *J. Biol. Chem.* 279, 51661–51668. doi:10.1074/jbc.M409325200.

Gonzalez-Garay, M. L. (2016). “Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq),” in (Springer, Dordrecht), 141–160. doi:10.1007/978-94-017-7450-5_6.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., and McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* 25, 1750–6. doi:10.1101/gr.191395.115.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–51. doi:10.1038/nrg.2016.49.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi:10.1126/science.1188021.

Hahne, F., and Ivanek, R. (2016). “Visualizing Genomic Data Using Gviz and Bioconductor,” in (Humana Press, New York, NY), 335–351. doi:10.1007/978-1-4939-3578-9_16.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project.

*Genome Res.* 22, 1760–74. doi:10.1101/gr.135350.111.

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489, 391–399. doi:10.1038/nature11405.

He, R., Ma, Y., Qi, Y., Jiang, S., Wang, N., Li, M., et al. (2012). Characterization of human cytomegalovirus UL146 transcripts. *Virus Res.* 163, 223–228. doi:10.1016/j.virusres.2011.09.034.

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., et al. (2018). A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* 19, 219. doi:10.1186/s12864-018-4611-3.

Heger, A. Pysam. Available at: github.com/pysam-developers/pysam [Accessed January 16, 2019].

Hussain, S., Aleksic, J., Blanco, S., Dietmann, S., and Frye, M. (2013). Characterizing 5-methylcytosine in the mammalian epitranscriptome. *Genome Biol.* 14, 215. doi:10.1186/gb4143.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062.

Isomura, H., Stinski, M. F., Kudoh, A., Murata, T., Nakayama, S., Sato, Y., et al. (2008). Noncanonical TATA sequence in the UL44 late promoter of human cytomegalovirus is required for the accumulation of late viral transcripts. *J. Virol.* 82, 1638–46. doi:10.1128/JVI.01917-07.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. doi:10.1038/nbt.4060.

Jones, E., Oliphant, T., and Peterson, P. (2001). {SciPy}: Open source scientific tools for {Python}.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* 96, 317–323. doi:10.1016/S1389-1723(03)90130-7.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–9.

doi:10.1093/bioinformatics/bts199.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., et al. (2006). CAGE: cap analysis of gene expression. *Nat. Methods* 3, 211–222. doi:10.1038/nmeth0306-211.

Kondo, K., Xu, J., and Mocarski, E. S. (1996). Human cytomegalovirus latent gene expression in granulocyte-macrophage progenitors in culture and in seropositive individuals. *Proc. Natl. Acad. Sci. U. S. A.* 93, 11137–42. doi:10.1073/pnas.93.20.11137.

Kronstad, L. M., Brulois, K. F., Jung, J. U., and Glaunsinger, B. A. (2013). Dual short upstream open reading frames control translation of a herpesviral polycistronic mRNA. *PLoS Pathog.* 9, e1003156. doi:10.1371/journal.ppat.1003156.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513, 409–413. doi:10.1038/nature13673.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–9. doi:10.1093/bioinformatics/btp352.

Li, X., Xiong, X., and Yi, C. (2017). Epitranscriptome sequencing technologies: decoding RNA modifications. *Nat. Methods* 14, 23–31. doi:10.1038/nmeth.4110.

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., et al. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* 359, 555–559. doi:10.1126/science.aao4426.

Ma, Y. P., Ruan, Q., Ji, Y. H., Wang, N., Li, M., Qi, Y., et al. (2011). Novel transcripts of human cytomegalovirus clinical strain found by cDNA library screening. *Genet. Mol. Res.* 10, 566–575. doi:10.4238/vol10-2gmr1059.

Ma, Y., Wang, N., Li, M., Gao, S., Wang, L., Zheng, B., et al. (2012). Human CMV transcripts: an overview. *Future Microbiol.* 7, 577–593. doi:10.2217/fmb.12.32.

Mader, R. M., Schmidt, W. M., Sedivy, R., Rizovski, B., Braun, J., Kalipciyan, M., et al. (2001). Reverse transcriptase template switching during reverse transcriptase–polymerase chain reaction: Artificial generation of deletions in ribonucleotide reductase mRNA. *J.*

*Lab. Clin. Med.* 137, 422–428. doi:10.1067/mlc.2001.115452.

Madoui, M.-A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., et al. (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* 16, 327. doi:10.1186/s12864-015-1519-z.

Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., et al. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* 11, 1455–1476. doi:10.1038/nprot.2016.086.

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159. doi:10.1038/nrg2521.

Moldován, N., Tombácz, D., Szűcs, A., Csabai, Z., Balázs, Z., Kis, E., et al. (2018). Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. *Sci. Rep.* 8, 8604. doi:10.1038/s41598-018-26955-8.

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., et al. (2017). The blood DNA virome in 8,000 humans. *PLOS Pathog.* 13, e1006292. doi:10.1371/journal.ppat.1006292.

Murphy, E., Rigoutsos, I., Shibuya, T., and Shenk, T. E. (2003). Reevaluation of human cytomegalovirus coding potential. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13585–90. doi:10.1073/pnas.1735466100.

Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., et al. (2017). Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell* 30, 149–161. doi:10.1007/s13577-017-0168-8.

Nam, D. K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., et al. (2002). Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6152–6. doi:10.1073/pnas.092140899.

Norris, A. L., Workman, R. E., Fan, Y., Eshleman, J. R., and Timp, W. (2016). Nanopore sequencing detects structural variants in cancer. *Cancer Biol. Ther.* 17, 246–253. doi:10.1080/15384047.2016.1139236.

Nyikó, T., Sonkoly, B., Mérai, Z., Benkovics, A. H., and Silhavy, D. (2009). Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. *Plant Mol. Biol.* 71, 367–378. doi:10.1007/s11103-009-9528-4.

O'Grady, T., Wang, X., Höner zu Bentrup, K., Baddoo, M., Concha, M., and Flemington, E. K. (2016). Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res.* 44, e145–e145. doi:10.1093/nar/gkw629.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., et al. (2009). Direct RNA sequencing. *Nature* 461, 814–818. doi:10.1038/nature08390.

Pérez Cañadillas, J. M., Varani, G., Neuhaus, D., and Varani, G. (2003). Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J.* 22, 2821–30. doi:10.1093/emboj/cdg259.

Potter, A. Analytical solutions for PacBio sequencing data. Available at: http://www.pacb.com/products-and-services/analytical-software/devnet/ [Accessed June 28, 2016].

Prazsák, I., Moldován, N., Balázs, Z., Tombácz, D., Megyeri, K., Szűcs, A., et al. (2018). Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* 19, 873. doi:10.1186/s12864-018-5267-8.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–2. doi:10.1093/bioinformatics/btq033.

Rahbar, A., Orrego, A., Peredo, I., Dzabic, M., Wolmer-Solberg, N., Stråat, K., et al. (2013). Human cytomegalovirus infection levels in glioblastoma multiforme are of prognostic value for survival. *J. Clin. Virol.* 57, 36–42. doi:10.1016/J.JCV.2012.12.018.

Rawlinson, W. D., and Barrell, B. G. (1993). Spliced transcripts of human cytomegalovirus. *J. Virol.* 67, 5502–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7688825 [Accessed July 17, 2017].

Rhoads, A., and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics. Proteomics Bioinformatics* 13, 278–89. doi:10.1016/j.gpb.2015.08.002.

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biol.* 14, 405. doi:10.1186/gb-2013-14-6-405.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–6. doi:10.1038/nbt.1754.

Roca, X., Sachidanandam, R., and Krainer, A. R. (2003). Intrinsic differences between authentic and cryptic 5' splice sites. *Nucleic Acids Res.* 31, 6321–33. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=275472&tool=pmcentrez&rendertype=abstract [Accessed February 4, 2016].

Roy, S. W., and Irimia, M. (2008). When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis. *BioEssays* 30, 601–605. doi:10.1002/bies.20749.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–5. doi:10.1126/science.1178534.

Schottstedt, V., Blümel, J., Burger, R., Drosten, C., Gröner, A., Gürtler, L., et al. (2010). Human Cytomegalovirus (HCMV) - Revised. *Transfus. Med. Hemother.* 37, 365–375. doi:10.1159/000322141.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–14. doi:10.1038/nbt.2705.

Shepard, P. J., Choi, E.-A., Lu, J., Flanagan, L. A., Hertel, K. J., and Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* 17, 761–72. doi:10.1261/rna.2581711.

Sheppard, S., Lawson, N. D., and Zhu, L. J. (2013). Accurate identification of polyadenylation sites from 3′ end deep sequencing using a naïve Bayes classifier. *Bioinformatics* 29, 2564–2571. doi:10.1093/bioinformatics/btt446.

Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* 34, 3955–3967. doi:10.1093/nar/gkl556.

Shnayder, M., Nachshon, A., Krishna, B., Poole, E., Boshkov, A., Binyamin, A., et al. (2018). Defining the Transcriptional Landscape during Cytomegalovirus Latency with Single-Cell RNA Sequencing. *MBio* 9, e00013-18. doi:10.1128/mBio.00013-18.

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* 4, 1521. doi:10.12688/f1000research.7563.2.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T. K., Hein, M. Y., Huang, S.-X., et al. (2012). Decoding human cytomegalovirus. *Science* 338, 1088–93. doi:10.1126/science.1227919.

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., et

al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* doi:10.1101/gr.222976.117.

Tian, B., Hu, J., Zhang, H., and Lutz, C. S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* 33, 201–212. doi:10.1093/nar/gki158.

Tombácz, D., Balázs, Z., Csabai, Z., Moldován, N., Szűcs, A., Sharon, D., et al. (2017a). Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. *Sci. Rep.* 7, 43751. doi:10.1038/srep43751.

Tombácz, D., Balázs, Z., Csabai, Z., Snyder, M., and Boldogkői, Z. (2018). Long-Read Sequencing Revealed an Extensive Transcript Complexity in Herpesviruses. *Front. Genet.* 9, 259. doi:10.3389/fgene.2018.00259.

Tombácz, D., Csabai, Z., Oláh, P., Balázs, Z., Likó, I., Zsigmond, L., et al. (2016). Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. *PLoS One* 11, e0162868. doi:10.1371/journal.pone.0162868.

Tombácz, D., Csabai, Z., Szűcs, A., Balázs, Z., Moldován, N., Sharon, D., et al. (2017b). Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1. *Front. Microbiol.* 8, 1079. doi:10.3389/fmicb.2017.01079.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi:10.1038/nprot.2012.016.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project. *Nature* 449, 804–810. doi:10.1038/nature06244.

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., et al. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* 361, k1687. doi:10.1136/BMJ.K1687.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends Genet.* 34, 666–681. doi:10.1016/J.TIG.2018.05.008.

Vancíková, Z., and Dvorák, P. (2001). Cytomegalovirus infection in immunocompetent and

immunocompromised individuals--a review. *Curr. Drug Targets. Immune. Endocr. Metabol. Disord.* 1, 179–87. doi:10.2174/1568005310101020179.

Vilela, C., and McCarthy, J. E. G. (2003). Regulation of fungal gene expression via short open reading frames in the mRNA 5'untranslated region. *Mol. Microbiol.* 49, 859–67. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12890013 [Accessed November 22, 2017].

Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., et al. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci.* 115, 9726–9731. doi:10.1073/PNAS.1806447115.

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7, 11708. doi:10.1038/ncomms11708.

Wang, S., Zhao, Z., Haque, F., and Guo, P. (2018). Engineering of protein nanopores for sequencing, chemical or protein sensing and disease diagnosis. *Curr. Opin. Biotechnol.* 51, 80–89. doi:10.1016/J.COPBIO.2017.11.006.

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., et al. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6, 100. doi:10.12688/f1000research.10571.2.

Wen, L. Z., Xing, W., Liu, L. Q., Ao, L. M., Chen, S. H., and Zeng, W. J. (2002). Cytomegalovirus infection in pregnancy. *Int. J. Gynecol. Obstet.* 79, 111–116. doi:10.1016/S0020-7292(02)00239-4.

Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip. Rev. RNA* 5, 765–768. doi:10.1002/wrna.1245.

Wickham, H. (2016). *Ggplot2 : elegrant graphics for data analysis*. New York: Springer-Verlag Available at: http://ggplot2.org [Accessed June 20, 2018].

Workman, R. E., Myrka, A. M., Wong, G. W., Tseng, E., Welch, K. C., and Timp, W. (2018). Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird Archilochus colubris. *Gigascience* 7. doi:10.1093/gigascience/giy009.

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–75. doi:10.1093/bioinformatics/bti310.

Yu, H., Wang, F., Tu, K., Xie, L., Li, Y.-Y., and Li, Y.-X. (2007). Transcript-level annotation of Affymetrix probesets improves the interpretation of gene expression data. *BMC Bioinformatics* 8, 194. doi:10.1186/1471-2105-8-194.

Yuan, C., Liu, Y., Yang, M., and Liao, D. J. (2013). New methods as alternative or corrective measures for the pitfalls and artifacts of reverse transcription and polymerase chain reactions (RT-PCR) in cloning chimeric or antisense-accompanied RNA. *RNA Biol.* 10, 1–11. doi:10.4161/rna.24570.

Zeng, X.-C., and Wang, S.-X. (2002). Evidence that BmTXK beta-BmKCT cDNA from Chinese scorpion Buthus martensii Karsch is an artifact generated in the reverse transcription process. *FEBS Lett.* 520, 183–4; author reply 185. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12044895 [Accessed August 15, 2018].

Zheng, D., Liu, X., and Tian, B. (2016). 3′READS+, a sensitive and accurate method for 3′ end sequencing of polyadenylated RNA. *RNA* 22, 1631–1639. doi:10.1261/rna.057075.116.

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., and Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30, 892–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11314272.

# 12 Copies of publications upon which the thesis was based